A Quasi Synthetic Control Method for Nonlinear Models With High-Dimensional Covariates *†‡

Zongwu Cai[a], Ying Fang[b,c], Ming Lin[b,c,†], Zixuan Wu[b]

[a]Department of Economics, University of Kansas, Lawrence, KS 66045, USA.
[b]Wang Yanan Institute of Studies in Economics, Xiamen University, Xiamen, Fujian 361005, China.
[c]Department of Statistics and Data Science, School of Economics, Xiamen University, Xiamen, Fujian 361005, China.

August 4, 2023

To make the conventional synthetic control method more flexible to estimate the average treatment effect, this article proposes a quasi synthesis control method for nonlinear models under the index model framework with possible high-dimensional covariates, together with a suggestion of using the minimum average variance estimation method to estimate parameters and the LASSO type procedure to choose covariates. Also, we derive the asymptotic distribution of the proposed estimators. A properly designed Bootstrap method is proposed to obtain confidence intervals and its theoretical justification is provided. Finally, Monte Carlo simulation studies are conducted to illustrate the finite sample performance and an empirical application to re-analyze the data from the National Supported Work Demonstration is also considered to demonstrate the proposed model to be practically useful.

Keywords: Average treatment effect; Bootstrap inference; Index model; Minimum average variance estimation method; Semiparametric estimation; Synthetic control method.

1 Introduction

When evaluating the impact of policy interventions, one of the main challenges lies in estimating unknown counterfactual outcomes. With observable covariates, a natural idea is

*The authors acknowledge the partially financial supports from the National Science Fund of China (NSFC) for Distinguished Scholars with the grant number 71625001, the NSFC key project with grant number 72033008, and the Science Foundation of Ministry of Education of China with the grant number 19YJA910003.

†*Corresponding author: Ming Lin*

‡E-mail addresses: caiz@ku.edu (Z. Cai), yifst1@xmu.edu.cn (Y. Fang), linming50@xmu.edu.cn (M. Lin), zixuan0801@stu.xmu.edu.cn (Z. Wu).

to construct an outcome regression model. In practice, the classic linear regression model is usually inadequate or even incorrect. To fully capture the relationship between the covariates and the outcomes, researchers suggest using the nonparametric model which can avoid the risk of model misspecification. However, the nonparametric model is challenged by the so-called *curse of dimensionality*. Therefore, as a combination of the parametric model and the nonparametric model, the semiparametric model has been conceived to overcome the aforementioned limitations.

There is a vast literature concerning applying semiparametric techniques to estimate the treatment effect and the existing research can be divided into two categories: estimating the counterfactual outcomes directly and indirectly. For the former, various semiparametric approaches have been used to estimate the conditional mean function or the conditional quantile function. For example, Heckman, Ichimura and Todd (1998) propose a kernel-matching-based estimator for the average treatment effect (ATE) and present a rigorous distributional theory, while Chiburis (2010) discusses the semiparametric bounds on the average treatment effect of a binary treatment on a binary outcome. Under the framework of the latent factor model to vary cross-section, Hsiao, Ching and Wan (2012) initiate an approach, termed as the panel data approach (PDA), that offers more flexibility than the DID, under a linear setting, and further, Ouyang and Peng (2015) extend the PDA approach to the nonlinear setting by allowing the conditional mean of the outcome to have a semiparametric form. As for the latter one, under the ignorability assumption, many scholars have proposed to first estimate the propensity score and then estimate the treatment effect of our interest by either re-weighting or matching technique. For details, see, for example, the papers by Abadie and Imbens (2006), Firpo (2007), Cattaneo (2010), Galvao and Wang (2015) and references therein.

One of the most important semiparametric models is the single index model. Friedman and Stuetzle (1981) propose the projection pursuit regression, which can be regarded as the prototype of the single index model. On the one hand, the single index model projects the multidimensional covariates into a one-dimensional single index variable by a linear transformation. On the other hand, it assumes an unknown nonlinear link function for the single index variable, which is greatly flexible. In the field of policy evaluation, the

single index model is usually used to analyze the relationship between the covariates and the treatment variables; see, instance, Park et al. (2021) and Sun, Yan and Li (2021). This relationship is measured by the so-called *propensity score*. The propensity score is unknown and needs to be estimated, which is sensitive to the model specification. As shown by Frölich (2004) and Kang and Schafer (2007), the misspecification of the propensity score can lead to misleading treatment effect estimates. Hence, using the single index model to flexibly characterize this relationship is advantageous.

In this article, our focus is on the statistical inference for the average treatment effect under the framework of the single index model. Our estimation procedure consists of two steps. First, parameters in the single index model are estimated by the minimum average variance estimation (MAVE) method proposed by Xia et al. (2002) and the penalized MAVE method by combining the bridge regression with MAVE, considered in Wang, Xu and Zhu (2013). In such a way, one can estimate the index with a possibility to take care of sparsity and choose covariates. In the second step, a nonparametric kernel smooth technique can be applied to estimate the weights for estimating the counterfactual outcomes. We make several contributions to the literature. First, our method is the first attempt to conduct a formal statistical inference for average treatment effects for single index models. Second, by using the generalized U-statistic technique for two samples, we derive the asymptotic inference theory for the corresponding ATE estimator. Third, we propose a properly designed (hybrid) Bootstrap method by combining the wild Bootstrap and the classical nonparametric Bootstrap and show that the carefully designed Bootstrap method provides valid inferences theoretically and empirically. Finally, we provide a simple sufficient condition under which the treatment effect estimator is uniquely determined and fast computing and show via simulations and an empirical example that the proposed method, which is robust to nonlinear model situations, can greatly enhance the applicability to estimating ATE. Therefore, our work complements the existing inference work in the literature on treatment effects.

The rest of the paper is organized as follows. Section 2 first presents the model setup for our method, and the estimation procedure is described in detail. Also, in this section, the asymptotic theory for the proposed estimator is given and a carefully designed Bootstrap method is provided with a theoretical justification for valid inferences. For choosing

covariates and taking care of sparsity, the penalized MAVE method is developed in the same section. A simulation study is conducted in Section 3 to illustrate the finite sample performance. Section 4 is devoted to reporting the empirical analysis of using our quasi synthetic control method to analyze the data from the National Supported Work (NSW) Demonstration. Finally, Section 5 concludes the paper. All detailed technical proofs are collected in Appendix.

2 Quasi Synthetic Control Method

2.1 Setup

Assume we observe n units and some of units are exposed to the treatment or intervention of our interest. The treatment status of unit i is indicated by a binary variable D_i, where $D_i = 1$ if i unit is treated and $D_i = 0$ otherwise. To define treatment effects, we adopt the potential outcomes framework proposed by Rubin (1974). Formally speaking, for each unit i, let Y_{1i} and Y_{0i} be the random variables representing potential outcomes under treatment and under no treatment, respectively. Then, the observed outcome Y_i can be written as $Y_i = D_i Y_{1i} + (1 - D_i)Y_{0i}$; that is to say, we can only observe Y_{1i} for the treated unit and Y_{0i} for the control unit. Besides, for each unit i, we can also observe a $(d \times 1)$ vector of pre-treatment predictors of Y_{0i} and denote it as $X_i \in \mathbb{R}^d$, where d might be relatively large. Assume there are n_1 units receiving the treatment and the remaining $n_0 = n - n_1$ units are not exposed to the treatment. For simplicity, we reorder these units so that the n_0 control units come first in the data set. Then, the observable data set can be written as $\{Y_i, D_i, X_i\}_{i=1}^n$ with $j = 1, \ldots, n_0$ being the control units and $i = n_0 + 1, \ldots, n$ being the treated units.

The quantities of our interest are the treatment effects at different levels. The basic one is the individual treatment effect on the treated units $i = n_0 + 1, \ldots, n$, which is defined as $\Delta_i = Y_{1i} - Y_{0i}$. Given Δ_i, the average treatment effect estimator is defined by:

$$\Delta = E(\Delta_i) = E(Y_{1i} - Y_{0i}). \tag{1}$$

The difficulty of estimating Δ_i and Δ lies in the fact that $\{Y_{0i}\}_{i=n_0+1}^n$ are not observable.

These unobservables are also named as the counterfactual outcomes and the estimation of the counterfactual outcomes constitutes the core of the research on the treatment effect.

2.2 Estimation Procedure

To consider a general setting, we consider the prediction function based on the conditional expectation of Y_{0i} given X_i, denoted by $m(x) = E(Y_{0i}|X_i = x)$, in an index form as $m(x) = m(\beta^\top x) = m(z)$, where $m(\cdot)$ is an unknown function and $z = \beta^\top x \in \mathbb{R}^1$, which covers the linear model as a special case. For the identification purpose, it is commonly assumed, in what follows, that the first element of β is positive and $||\beta||^2 = \sum_{k=1}^d \beta_k^2 = 1$. Then, for $i = n_0 + 1, \ldots, n$, $E(Y_{0i}) = E\left[E(Y_{0i}|X_i)\right] = E\left[E(Y_{0i}|Z_i)\right]$, where $Z_i = \beta^\top X_i$ for a given β, so that the estimation of $m(z)$ is one-dimensional and the so-called *curse of dimensionality* in a nonparametric smoothing can be avoided. Under Assumption A1 in Section 2.3, the kernel type (Nadaraya-Watson)[1] estimate of $m(z)$, based on the data $\{(Y_j, X_j)\}_{j=1}^{n_0}$ from pre-intervention, is given by

$$\tilde{m}(z) = \sum_{j=1}^{n_0} c_{j,h}(z) Y_j, \tag{2}$$

where $c_{j,h}(z) = K_h(Z_j - z)/\sum_{l=1}^{n_0} K_h(Z_l - z)$, $K_h(u) = K(u/h)/h$, and $K(u)$ is a kernel function, and h is the bandwidth. Now, the infeasible prediction of Y_{0i} is denoted by $\tilde{Y}_{0i}$

$$\tilde{Y}_{0i} = \tilde{m}(Z_i) = \sum_{j=1}^{n_0} c_{j,h}(Z_i) Y_j \tag{3}$$

for $i = n_0 + 1, \ldots, n$. Actually, (3) is infeasible since it is based on the unknown quantities $\{Z_j\}_{j=1}^{n_0}$. Accordingly, the infeasible estimate of Δ, $\tilde{\Delta}$ is given by

$$\tilde{\Delta} = \frac{1}{n_1} \sum_{i=n_0+1}^{n} \left[Y_{1i} - \sum_{j=1}^{n_0} c_{j,h}(Z_i) Y_j \right] = \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i - \frac{1}{n_0} \sum_{j=1}^{n_0} a_{j,h} Y_j, \tag{4}$$

where $a_{j,h} = a_h(Z_j)$ and

$$a_h(z) = \frac{1}{n_1} \sum_{i=n_0+1}^{n} K_h(Z_i - z) \left[\frac{1}{n_0} \sum_{l=1}^{n_0} K_h(Z_l - Z_i) \right]^{-1}.$$

[1] Of course, one can use the kernel smoothing technique such as the local polynomial estimation method as in Fan and Gijbels (1996).

Clearly, under this nonlinear setting, we need to find the weights β such that $\beta^\top X_i$ can be the best to predict Y_{0i} for $i = 1, \ldots, n_0$. In other words, we need to search for β such as its conditional expectation of Y_{0i} given Z_i matches the conditional expectation of Y_{0i} given X_i as close as possible. Therefore, to do so, we suggest using the index model and its estimation approach described in Section 2.4.

Interestingly, our method shares some similarities and differences with the synthetic control method (SCM) proposed by Abadie and Gardeazabal (2003), which has been described as "arguably the most important innovation in the policy evaluation literature in the last 15 years" as argued by Athey and Imbens (2017). Although the SCM is originally designed to deal with the panel data setting, Abadie and L'Hour (2021) present a penalized version of the SCM for disaggregated data. For each treated unit $i = n_0 + 1, \ldots, n$, a synthetic control can be represented by a $n_0 \times 1$ vector of weights, $\mathbf{W}_i^{sc} = (W_{i,1}^{sc}, \ldots, W_{i,n_0}^{sc})^\top$. Given a set of weights, $\mathbf{W}_i^{sc}$, the synthetic control estimators (linear predictor) of Y_{0i} and Δ_i are $\hat{Y}_{0i} = \sum_{j=1}^{n_0} W_{i,j}^{sc} Y_j$ and $\hat{\Delta}_i = Y_{1i} - \hat{Y}_{0i}$, respectively. Apparently, $c_{j,h}(Z_i)$ in (3) is identical to the SCM weights $\mathbf{W}_i^{sc}$ defined in Equation (4) in Abadie and L'Hour (2021). However, different from Abadie and L'Hour (2021), our weights $\{c_{j,h}(Z_i)\}$ takes care of both the best prediction to resemble the characteristics of the treated unit before the intervention and nonlinearity of prediction function since our model is in a semiparametric nature. Besides, our approach does not require that weights $a_{j,h}$ should satisfy the standard constraints as in Abadie and L'Hour (2021). Instead, our method is similar to that for the PDA as in Hsiao, Ching and Wan (2012) and Wan, Xie and Hsiao (2018) as well as Ouyang and Peng (2015), in the sense that it does not have constraints on weights such as nonnegative weights. Therefore, our method is termed as the quasi synthetic control method (QSCM), although both have different motivations.

Finally, from the above discussions, the QSCM estimation procedure for estimating Δ consists of the following two steps. First, use (9) given in Section 2.4 to obtain $\hat{\beta}$, and then, set $\hat{Z}_j = \hat{\beta}^\top X_j$ for $j = 1, \cdots, n_0$ and $\hat{Z}_i = \hat{\beta}^\top X_i$ for $i = n_0 + 1, \ldots, n$. Second, compute the

feasible estimate of Δ based on (4), and $\hat{\Delta}$ is defined as

$$\hat{\Delta} = \frac{1}{n_1} \sum_{i=n_0+1}^{n} \left[Y_{1i} - \sum_{j=1}^{n_0} \hat{c}_{j,h}(\hat{Z}_i) Y_j \right] = \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i - \frac{1}{n_0} \sum_{j=1}^{n_0} \hat{a}_{j,h} Y_j, \tag{5}$$

where $\hat{a}_{j,h} = \hat{a}_h(\hat{Z}_j) = \frac{1}{n_1} \sum_{i=n_0+1}^{n} K_h(\hat{Z}_i - \hat{Z}_j) \left[\frac{1}{n_0} \sum_{l=1}^{n_0} K_h(\hat{Z}_l - \hat{Z}_i) \right]^{-1}$, which is similar to $W_j^{sc} = (n_0/n_1) \sum_{i=1}^{n_1} W_{i,j}^{sc}$ as in Equation (4) in Abadie and L'Hour (2021).

2.3 Asymptotic Theory

To describe the asymptotic properties of $\hat{\Delta}$, some notations are introduced. Let $f_c(z)$ be the density of Z_j for $j = 1, \ldots, n_0$ and $f_t(z)$ be the density of Z_i for $i = n_0 + 1, \ldots, n$. Define $\mathcal{C}_1$ to be the support of Z_j for $j = 1, \ldots, n_0$ and $\mathcal{C}_2$ to be the support of Z_i for $i = n_0 + 1, \ldots, n$. Define the CDF of Y_{1i}, $F_{Y_{1i}}(\cdot)$ and its density function $F'_{Y_{1i}}(y) = f_{Y_{1i}}(y)$.

Assumptions:

A1. Assume that the conditional expectation of outcome Y_{0j} given predictor X_j for $j = 1, \ldots, n_0$, denoted by $m_p(x)$, is the same as the conditional expectation of outcome Y_{0i} given predictor X_i for $i = n_0 + 1, \ldots, n$, denoted by $m_a(x)$; that is, $m_p(x) = m_a(x) = m(x)$. Also, assume that $m(x)$ is in the form of index $z = \beta^\top x$; that is, $m(x) = m(z)$. Furthermore, assume that the second order derivative of $m(z)$ is continuous. Finally, assume that the first element of β is positive and $||\beta||^2 = \sum_{k=1}^{d} \beta_k^2 = 1$.

A2. $\{Y_{0i}, Y_{1i}, X_i\}_{i=1}^{n}$ are independent and identically distributed. Assume that $E(|Y_i|^s) < \infty$ for some $s > 2$, $\mathcal{C}_2 \subseteq \mathcal{C}_1$, and $f_t(z) \geq M_1 > 0$. Also, assume that $F_{Y_{1i}}(\cdot)$ is twice differentiable and $f_{Y_{1i}}(\cdot) > 0$.

A3. Assume that $0 < \lambda < \infty$, where $\lim n_1/n_0 = \lambda$, and $n_0 h^2 \to \infty$ and $n_0 h^4 \to 0$ as $n_0 \to \infty$.

A4. The kernel function $K(\cdot)$ is symmetric and is bounded positive function as well as satisfies a Lipschitz condition. And the first derivative of $K(\cdot)$ is continuous.

A5. Assume that the second order of derivative of $r(z)$ is bounded, where $r(z) = f_t(z)/f_c(z)$, the ratio function to characterize the distributional changes of the single index between the treated units and control units.

A6. Assume that for any estimate of β, $\hat{\beta}$ admits the following expression

$$\sqrt{n_0}\left[\hat{\beta}-\beta\right]=\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\phi(X_j,Y_j)+o_p(1)\ \rightarrow\ N(0,\Sigma_\beta) \tag{6}$$

for some function $\phi(\cdot)$ with the finite variance Σ_β =Var$(\phi(X_j,Y_j))$ for $j=1,\ldots,n_0$.

Assumptions listed above are standard. In particular, Assumption A1 is to assume that the conditional expectations before and after treatments are same, which is also imposed in Hsiao, Ching and Wan (2012). Assumption A3 is under-smoothed in a nonparametric kernel smoothing estimation, which makes the asymptotic bias negligible. Also, this assumption leads the practical choice of h in application to be not difficult. Clearly, $r(z)=1$ in Assumption A5 if $f_c(z)=f_t(z)$, which means that there is no distributional change of the single index bewteen the treated units and control units. Indeed, $r(z)$ is interpreted as "acceptance probability" in rejection sampling instead of "importance re-weighting", or *covariate shift*, in the machine learning literature; see, for example, Wu, Ren and Mu (2016) and references therein. The assumption in A6 is common in the index model literature; see, for example, Cai, Juhl and Yang (2015) and references therein. Indeed, under some regularity conditions, $\sqrt{n_0}\left[\hat{\beta}-\beta\right]$ can be expressed as in (6), which holds true; see, for instance, Xia (2006) or Section 2.4 for details. Finally, define $\sigma_3^2=\delta_a^\top\Sigma_\beta\delta_a$, where $\delta_a=E\left[m'(Z_i)X_i^\top\right]$ for $i=n_0+1,\ldots,n$, where $m'(z)$ is the first order derivative of $m(z)$, and Σ_β is given in Assumption A6. Also, define Σ_{23} =Cov$(\phi(X_j,Y_j),r(Z_j)\varepsilon_j)$, where $\varepsilon_j=Y_{0j}-E(Y_{0j}\,|\,X_j)$ for $j=1,\ldots,n_0$. Now, the asymptotic normality of $\hat{\Delta}$ is stated in the following theorem with its theoretical proof relegated to Appendix, based on the generalized U-statistic theory for two samples given in Serfling (1980, p.175).

Theorem 1: Under Assumptions A1 - A6, one has

$$\sqrt{n_1}\left[\hat{\Delta}-\Delta\right]\ \rightarrow\ N(0,\sigma_\Delta^2),$$

where $\sigma_\Delta^2=\sigma_1^2+\lambda\left[\sigma_2^2+\sigma_3^2+2\,\delta_a^\top\Sigma_{23}\right]$ with $\sigma_1^2=\mathrm{Var}\left(Y_{1i}-m(Z_i)\right)$ for $i=n_0+1,\ldots,n$ and $\sigma_2^2=\mathrm{Var}\{r(Z_j)\varepsilon_j\}$ for $j=1,\ldots,n_0$.

It follows from Theorem 1 that the asymptotic variance consists of four terms. In particular, the first term in σ_Δ^2 stands for the variance of $Y_{1i}-m(Z_i)$, which is the same as the

variance of $\Delta_i + \varepsilon_i$, the second term characterizes the variation for estimating Y_{0i}, the third term σ_3^2 is the variation carried over from the estimation of β, and the last term depicts the correlation between the first step and the second step. This is typical for a two-stage procedure as addressed in Cai et al. (2006). Also, one can see that obtaining a consistent estimate of σ_Δ^2 is not a straight forward task due to its complicated form of involving several terms. However, a Bootstrap procedure can overcome this difficulty; see Section 2.5 for details.

2.4 MAVE Method for Estimating β

Now, it turns to discussing how to estimate β. To do so, let Y be a random variable and $X = (X_1, \cdots, X_d)^\top$ be a collection of d random variables. The single index model, one of the most popular semiparametric models in statistica and econometrics, can be written as

$$Y = m(\beta^\top X) + \varepsilon = m(Z) + \varepsilon, \tag{7}$$

where $E(\varepsilon|X) = 0$, $m(\cdot)$ is an unknown link function, and $\beta = (\beta_1, \cdots, \beta_d)^\top$ is the $d \times 1$ index vector. If $m(v) = v$, the model in (7) reduces to the linear model in (5), so that our model is in a semiparametric nature. For the sake of identification, it is usually assumed that $\beta_1 = 1$ or $\beta^\top\beta = 1$ with $\beta_1 > 0$. From (7), one can see that the linear combination $Z = \beta^\top X = \beta_1 X_1 + \cdots + \beta_d X_d$ captures all the information of X on Y. The estimation of the index vector β has attracted extensive attentions. For example, Ichimura (1993) proposes the semiparametric least squares estimation of the single index model based on the leave-one-out technique. Due to the fact that the single index model shares a close connection with the central mean subspace in the sufficient dimension reduction, Xia et al. (2002) propose the (conditional) minimum average variance estimation method for the dimension reduction problem and later, Xia (2006) shows that this method can be applied to the single index model. Therefore, the MAVE method proposed in Xia (2006) is employed in our setting to estimate β, described as follows.

Notice that under the least squares loss,

$$\beta = \arg\min_{\tilde{\beta} \in \mathbb{R}^d} E\left[Y - E(Y|\tilde{\beta}^\top X)\right]^2. \tag{8}$$

In our setting, the index is estimated by the observed data for the control units, $\{Y_j,\, X_j\}_{j=1}^{n_0}$. Motivated by the local linear smoothing technique, the sample analogue of (8) is given by

$$\beta = \arg \min_{\substack{\tilde{\beta}\in\mathbb{R}^d:\tilde{\beta}^\top\tilde{\beta}=1 \\ a_j,b_j}} \sum_{j=1}^{n_0}\sum_{i=1}^{n_0} \left[Y_i - a_j - b_j\tilde{\beta}^\top(X_i - X_j)\right]^2 w_{ij}, \tag{9}$$

where $w_{ij} = K_{h_0}(\tilde{\beta}^\top(X_i - X_j))$ with $K_{h_0}(v) = K(v/h_0)/h_0$ and $K(\cdot)$ being a kernel function as well as h_0 being the bandwidth. Define $X_{ij} = X_i - X_j$, Xia (2006) proposes the following algorithm for estimating β:

Step 1. Set an initial value $\beta^{(0)}$.

Step 2. For $k \geq 1$, calculate

$$\begin{pmatrix} \hat{a}_j^{\beta^{(k-1)}} \\ \hat{b}_j^{\beta^{(k-1)}} h_0 \end{pmatrix} = \left\{\sum_{j=1}^{n_0} K_{h_0}\left(\beta^{(k-1)^\top} X_{ij}\right) Z_{ij}^{(k-1)} Z_{ij}^{(k-1)^\top}\right\}^{-1} \sum_{j=1}^{n_0} K_{h_0}\left(\beta^{(k-1)^\top} X_{ij}\right) Z_{ij}^{(k-1)} Y_j,$$

where $Z_{ij}^{(k-1)} = \left(1, \beta^{(k-1)^\top} X_{ij}/h_0\right)^\top$, and also, obtain

$$\hat{f}_{\beta^{(k-1)}}(\beta^{(k-1)^\top} X_j) = \frac{1}{n_0}\sum_{i=1}^{n_0} K_{h_0}(\beta^{(k-1)^\top} X_{ij}), \quad \text{and} \quad \hat{\rho}_j^{\beta^{(k-1)^\top}} = \rho_{n_0}(\hat{f}_{\beta^{(k-1)}}(\beta^{(k-1)^\top} X_j)),$$

where $\rho_{n_0}(\cdot)$ is a trimming function for the boundary points. Following the suggestion from Xia (2006), $\rho_{n_0}(v)$ is chosen as a bounded function with bounded derivative on $\mathbb{R}$ such that $\rho_{n_0}(v) = I(v > 2c_0 n_0^{-\varepsilon})$, where $I(A)$ is the indicator function of set A.

Step 3. Calculate

$$\begin{aligned}\beta^{(k)} = &\left\{\sum_{i=1}^{n_0}\sum_{j=1}^{n_0} K_{h_0}\left(\beta^{(k-1)^\top} X_{ij}\right) \hat{\rho}_j^{\beta^{(k-1)}} \left(\hat{b}_j^{\beta^{(k-1)}}\right)^2 X_{ij} X_{ij}^\top / \hat{f}_{\beta^{(k-1)}}\left(\beta^{(k-1)^\top} X_j\right)\right\}^{-1} \\ &\times \sum_{i=1}^{n_0}\sum_{j=1}^{n_0} K_{h_0}\left(\beta^{(k-1)^\top} X_{ij}\right) \hat{\rho}_j^{\beta^{(k-1)}} \hat{b}_j^{\beta^{(k-1)}} X_{ij}\left(Y_i - \hat{a}_j^{\beta^{(k-1)}}\right) / \hat{f}_{\beta^{(k-1)}}\left(\beta^{(k-1)^\top} X_j\right).\end{aligned}$$

Step 4. Set $\beta^{(k)} = \text{sign}(\beta_1^{(k)})\beta^{(k)}/\|\beta^{(k)}\|$. Then, repeat Steps 2 and 3 until convergence reaches.

Denote the ultimate estimator for β as $\hat{\beta}_{\text{MAVE}}$. Theoretically, Xia (2006) derives the asymptotic normality for $\hat{\beta}_{\text{MAVE}}$ and shows that the asymptotic covariance matrix of $\hat{\beta}_{\text{MAVE}}$ can achieve the information lower bound in the semiparametric sense. From Xia (2006),

one can see that under some regularity conditions, $\hat{\beta}_{\text{MAVE}}$ satisfies (6) with $\phi(X_j, Y_j) = W_{m_0}^{+} m'(\beta^\top X_j) v_\beta(X_j)\varepsilon_j$, where $m'(z)$ is the first derivative of $m(z)$, $v_\beta(x) = E(X|\beta^\top X = \beta^\top x) - x$, $W_{m_0} = E\{m'(\beta^\top X)^2 v_\beta(X) v_\beta^\top(X)\}$, and $W_{m_0}^{+}$ is the Moore-Penrose inverse of W_{m_0}, while its asymptotic variance is given as $\Sigma_\beta = [E\{m'(\beta^\top X)^2 W(X)\}]^{+} E\{m'(\beta^\top X)^2 W_0(X)\varepsilon^2\}$ $[E\{m'(\beta^\top X)^2 W(X)\}]^{+}$, where $W_0(x) = v_\beta(x) v_\beta^\top(x)$ and $W(x) = E(XX^\top|\beta^\top X = \beta^\top x) - E(X|\beta^\top X = \beta^\top x) E^\top(X|\beta^\top X = \beta^\top x)$. Therefore, the assumption in Assumption A6 is not a big concern.

2.5 A Bootstrap Inference

Clearly, Theorem 1 provides the asymptotic distribution for $\hat{\Delta}$, so that an inference can be made if σ_Δ^2 can be estimated consistently. But, one can see from Theorem 1 that the form of σ_Δ^2 is complicated so that it is not easy to get a consistent estimate. Therefore, it is a difficult task to construct a confidence interval (CI) for Δ. To facilitate an easy inference, we propose the following (hybrid) Bootstrap procedure by combining the (conditional) wild Bootstrap similar to that in Zhang, Huang and Liu (2020) for single index models and the nonparametric Bootstrap, to estimate σ_Δ^2.

Step 1. Given $\{Y_j, X_j\}_{j=1}^{n_0}$ and $\{Y_i, X_i\}_{i=n_0+1}^{n}$, estimate the treatment effect by (5) as $\hat{\Delta}$.

Step 2. Generate the wild Bootstrap sample $\{(X_j, Y_j^*)\}_{j=1}^{n_0}$ of the control group, where for $1 \le j \le n_0$, $Y_j^* = \hat{Y}_j + \varepsilon_j^*$ with $\varepsilon_j^* = (Y_j - \hat{Y}_j)\xi_j$ and $\{\xi_j\}_{j=1}^{n_0}$ are i.i.d. random errors with mean zero and unit variance conditional on the original sample $\{X_j, Y_j\}_{j=1}^{n_0}$.

Step 3. Generate the nonparametric Bootstrap sample $\{(X_i^*, Y_i^*)\}_{i=n_0+1}^{n}$ of the treated group by drawing with replacement from the original dataset $\{(X_i, Y_i)\}_{i=n_0+1}^{n}$.

Step 4. Using the wild Bootstrap sample $\{(X_j, Y_j^*)\}_{j=1}^{n_0}$ to re-estimate the index parameter as $\hat{\beta}^*$. Set $\hat{Z}_j^* = X_j^\top \hat{\beta}^*$ for $j = 1, \ldots, n_0$ and $\hat{Z}_i^* = (X_i^*)^\top \hat{\beta}^*$ for $i = n_0 + 1, \ldots, n$. Then, obtain the quasi synthetic control estimator $\hat{\Delta}^*$ as

$$\hat{\Delta}^* = \frac{1}{n_1} \sum_{i=n_0+1}^{n} \left[Y_i^* - \sum_{j=1}^{n_0} \hat{c}_{j,h}^*(\hat{Z}_i^*) Y_j^* \right] = \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i^* - \frac{1}{n_0} \sum_{j=1}^{n_0} \hat{a}_{j,h}^* Y_j^*,$$

where $\hat{a}_{j,h}^* = \hat{a}_h^*(\hat{Z}_j^*) = \frac{1}{n_1} \sum_{i=n_0+1}^{n} K_h(\hat{Z}_i^* - \hat{Z}_j^*) \left[\frac{1}{n_0} \sum_{l=1}^{n_0} K_h(\hat{Z}_l^* - \hat{Z}_i^*) \right]^{-1}$, which is the Bootstrap version of $\hat{a}_{j,h}$ in (5).

Step 5. Repeat steps 2 to 4 a large number of times, say, B times to obtain $\{\hat{\Delta}^{*(b)}\}_{b=1}^{B}$. Then σ_{Δ}^2 can be estimated as $n_1 \sum_{b=1}^{B}(\hat{\Delta}^{*(b)} - \hat{\Delta})^2/(B-1)$, denoted as $\hat{\sigma}_{\Delta}^2$.

Note that the reason on proposing the above hybrid Bootstrap is that at Step 2, the wild Bootstrap is used since both $m(\cdot)$ and β in (7) need to be re-estimated by the Bootstrap sample of the control group and at Step 3, a simple nonparametric Bootstrap is employed since there is no re-estimation involving the Bootstrap sample of the treated group. Finally, a (1-α)100% Bootstrap CI for Δ can be constructed as $\hat{\Delta} \pm z_{\alpha/2}\hat{\sigma}_{\Delta}/\sqrt{n_1}$ based on the asymptotic normality of $\hat{\Delta}$ in Theorem 1, where $z_{\alpha/2}$ is the $(1-\alpha/2)$th percentile of the standard normal distribution. The theoretical validity of this Bootstrap procedure can be confirmed by the following theorem with its detailed proof presented in Appendix, and in Section 3, it is empirically illustrated by simulation to demonstrate the finite sample performance of the proposed Bootstrap procedure.

Theorem 2: Under the conditions imposed in Theorem 1, conditional on the original sample $\{X_j, Y_j\}_{j=1}^{n_0}$ and $\{X_i, Y_i\}_{i=n_0+1}^{n}$ and in probability, one has

$$\sqrt{n_1}\left(\hat{\Delta}^* - \hat{\Delta}\right) \xrightarrow{d} \mathrm{N}(0, \sigma_{\Delta}^2),$$

where σ_{Δ}^2 is defined in Theorem 1.

Alternatively, one might apply the subsampling technique as proposed in Li (2020) to construct the CI for Δ. Li (2020) suggests to decompose the SCM-oriented ATE estimator $\hat{\Delta}$ into two terms as follows. The first term is related to the synthetic control weight estimator, and the second term is unrelated to the weight estimator but depends on the sample size of the treated units n_1. The weight estimator is constrained so that the standard Bootstrap method may be inconsistent when the true parameter lies on the boundary of the parameter space as addressed in Li (2020), while the subsampling method is not distracted by this constraint. Therefore, Li (2020) proposes a subsampling-Bootstrap method for the inference of the synthetic control method, which applied the subsampling method to the term related to the constrained estimator and applied the Bootstrap method to the remaining term.

Besides, our method shares a deep connection with the popular matching methods. Actually, Abadie and Imbens (2011) demonstrate that the standard Bootstrap method fails

to conduct inference for matching estimators. To overcome this problem, Otsu and Rai (2017) propose asymptotically valid inference methods for matching estimators based on the weighted Bootstrap. However, their method only deals with the case of a fixed number of matches. Our method matches each treated unit with all control units, which means that the number of matches increases with the size of the control group and is definitely not fixed.

2.6 Choosing Covariates

Based on the above discussion, we actually assume a single index model, as in (7), for Y_{0i}. For the single index model, when the number of predictor variables is large, it is necessary to discriminate relevant variables from irrelevant variables, since the inclusion of irrelevant variables may harm estimation accuracy and model interpretability. This negative effect of including irrelevant variables may be amplified in our quasi synthetic control method due to the fact that our method is intrinsically a two step procedure.

Many classical variable selection procedures have been generalized to the single index model. For example, Naik and Tsai (2001) derive a bias-corrected version of Akaike's information criterion, AIC_c, for the single index model that selects not only relevant variables but also a smoothing parameter for the unknown link function, while Kong and Xia (2007) propose the separation validation technique for the variable selection in the single index model.

From the discussion in Section 2.4, we know that the MAVE estimate of β is obtained by solving the minimization problem (9). Generally, to select the relevant variables, we can add a penalty term to the least-squares-form loss function in (9):

$$\sum_{j=1}^{n_0}\sum_{i=1}^{n_0}[Y_i - a_j - b_j\tilde{\beta}^\top(X_i - X_j)]^2 w_{ij} + n_0\sum_{k=1}^{d} p_{\lambda_{n_0}}(|\tilde{\beta}_k|),$$

where $p_\lambda(\cdot)$ denotes a penalty function and λ_{n_0} denotes the penalty parameter. Different choices of $p_\lambda(\cdot)$ can lead to different variable selection methods.

The simplest choice is to set $p_{\lambda_{n_0}}(|\tilde{\beta}_j|) = \lambda_{n_0}|\tilde{\beta}_j|$, which corresponds to the well-known least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996). Indeed, Wang and Yin (2008) adopt this L_1 norm penalty and proposed the sparse MAVE

method and Zeng, He and Zhu (2012) further explore the idea of combining MAVE and LASSO, and propose the sim-LASSO method. The sim-LASSO method not only penalizes the L_1 norm of the index parameter β, but also penalizes the terms $\{b_j\}_{j=1}^{n_0}$ in (9). Since $b_j = \partial m(u)/\partial u|_{u=\beta^\top X_j}$, adding this penalty contributes to excluding the data points with less information on estimating β, which stabilizes and improves the estimation of β. Finally, Wang, Xu and Zhu (2013) propose the penalized MAVE method by combining the bridge regression with MAVE. In the case of the single-index-model, the penalized MAVE estimator has the oracle property.

It is widely accepted that a good penalty function should lead to an unbiased, sparse and continuous estimator. However, the LASSO estimator is biased for large parameters. Alternatively, Fan and Li (2001) propose the smoothly clipped absolute deviation (SCAD) penalty. The SCAD penalty is defined via its first derivative as $p'_\lambda(\beta_k) = \lambda I(\beta_k \leq \lambda) + (a\lambda - \beta_k)_+ I(\beta_k > \lambda)/(a-1)$ for some $a > 2$. Due to the oracle property of the SCAD penalty justified by Fan and Li (2001), Peng and Huang (2011) explore the idea of introducing the SCAD penalty into the single index model. Given that the dimension of β is a fixed constant, the SCAD estimator has the oracle property. Hence, we can also combine the SCAD penalty with MAVE, and modify the objective function in (9) as:

$$\beta = \arg \min_{\substack{\tilde{\beta}\in\mathbb{R}^d:\tilde{\beta}^\top\tilde{\beta}=1 \\ a_j,b_j}} \left\{ \sum_{j=1}^{n_0}\sum_{i=1}^{n_0} \left[Y_i - a_j - b_j\tilde{\beta}^\top(X_i - X_j) \right]^2 w_{ij} + n_0 \sum_{k=1}^{d} p^{\mathrm{SCAD}}_{\lambda_{n_0}}(|\tilde{\beta}_k|) \right\}, \quad (10)$$

where $w_{ij} = K_{h_1}(\tilde{\beta}^\top(X_i - X_j))$ with $K_{h_1}(v) = K(v/h_1)/h_1$ and $K(\cdot)$ being a kernel function as well as h_1 being the bandwidth. Similarly, the optimization problem in (10) can be solved alternatively and iteratively, and the SCAD-MAVE algorithm can be summarized as follows:

Step 1. Given data $\{Y_j, X_j\}_{j=1}^{n_0}$, calculate the initial estimator $\hat{\beta}_0$ by the MAVE method. Set $t = 1$.

Step 2. Given $\hat{\beta}_{(t-1)}$, calculate the refined weights as

$$w_{ij}^{(t-1)} = K_{h_1}\left[\hat{\beta}_{(t-1)}^\top(X_i - X_j)\right] \Big/ \sum_{l=1}^{n_0} K_{h_1}\left[\hat{\beta}_{t-1}^\top(X_l - X_j)\right].$$

Then, solve the inner optimization problem for $j = 1, \ldots, n_0$:

$$\min_{a_j, b_j} \sum_{i=1}^{n_0} \left[Y_i - a_j - b_j \hat{\beta}_{(t-1)}^{\top} (X_i - X_j) \right]^2 w_{ij}^{(t-1)}$$

Clearly, this problem is analogous to the weighted least squares problem. We can easily derive the analytical solutions and denote them as $\hat{a}_j^{(t-1)}$ and $\hat{b}_j^{(t-1)}$.

Step 3. Given $\hat{a}_j^{(t-1)}$ and $\hat{b}_j^{(t-1)}$, we solve the outer optimization problem:

$$\min_{\tilde{\beta} \in \mathbb{R}^d : \tilde{\beta}^{\top} \tilde{\beta} = 1} \left\{ \sum_{j=1}^{n_0} \sum_{i=1}^{n_0} \left[Y_i - \hat{a}_j^{(t-1)} - \hat{b}_j^{(t-1)} \tilde{\beta}^{\top} (X_i - X_j) \right]^2 w_{ij} + n_0 \sum_{k=1}^{d} p_{\lambda_{n_0}}^{\text{SCAD}}(|\tilde{\beta}_k|) \right\}$$

Obviously, regardless of the constraint $\tilde{\beta}^{\top} \tilde{\beta} = 1$, we can rewrite the first part in least squares form, then we can use the *ncvreg* package in R to optimize it and obtain the estimator $\hat{\beta}^{(t)}$. Let $\hat{\beta}^{(t)} = \text{sign}(\hat{\beta}_1^{(t)}) \hat{\beta}^{(t)} / \|\hat{\beta}^{(t)}\|$.

Step 4. Check whether $\|\hat{\beta}^{(t)} - \hat{\beta}^{(t-1)}\|^2 < c$, where c is an arbitrarily small positive constant, if not, set $t = t + 1$ and go to Step 2. Denote the final estimator as $\hat{\beta}_{\text{SCAD-MAVE}}$.

Based on the above discussion, we can use the SCAD-MAVE method to select relevant variables at first, then, set $\hat{Z}_j = \hat{\beta}_{\text{SCAD-MAVE}}^{\top} X_j$ for $j = 1, \ldots, n_0$ and $\hat{Z}_i = \hat{\beta}_{\text{SCAD-MAVE}}^{\top} X_i$ for $i = n_0 + 1, \ldots, n$. Finally, we can estimate the treatment effect by (5), denoted by $\hat{\Delta}_{\text{SCAD}}$. To derive the asymptotic property of $\hat{\Delta}_{\text{SCAD}}$, we make following assumptions.

B1. For $l = 1, \ldots, n$, $Y_{0l} = m(\beta^{\top} X_l) + \varepsilon_l$, where $E(\varepsilon_l | X_l) = 0$, $E(\varepsilon_l^2 | X_l) = \sigma^2 > 0$, and $E(\varepsilon_l^4 | X_l)$ exists.

B2. The marginal density of $X^{\top} \tilde{\beta}$ is positive and uniformly continuous in a neighborhood of β. Furthermore, $X^{\top} \tilde{\beta}$ has a positive density on its support.

B3. Assume that the density function of X has a continuous second derivative and $\text{Cov}(X)$ is nonsingular. Furthermore, assume that for any vector v, if $v^{\top} \text{Cov}(X) \beta = 0$, then $v^{\top} \Sigma v > c||v||^2$, where $\Sigma = E\left\{ [m'(Z)]^2 \left[X - E\left(X | Z = \beta^{\top} X\right) \right] \left[X - E\left(X | Z = \beta^{\top} X\right) \right]^{\top} \right\}$.

B4. $n_0 \, h_1^3 \to \infty$ and $n_0 \, h_1^4 \to 0$ as $n_0 \to \infty$.

On can see that the above assumptions are indeed regularity assumptions, also listed in Peng and Huang (2011). Without loss of generality, we assume that the first s components of β are non-zeros, i.e. β is partitioned as $\beta_{\mathcal{A}} = (\beta_1, \ldots, \beta_s)^{\top}$ and $\beta_{\mathcal{A}^C} = (0, \ldots, 0)^{\top}$ with $d - s$ components. Under these assumptions, we can conclude that $\hat{\beta}_{\text{SCAD-MAVE}}$ satisfies (6)

by Theorem 2 in Peng and Huang (2011).

Lemma 1: Under Assumptions B1 - B4, if the tuning parameter λ_{n_0} satisfies $\lambda_{n_0} \to 0$ and $\sqrt{n_0}\lambda_{n_0} \to \infty$, then, with probability approaching 1, we have:

(a) Sparsity: $\hat{\beta}_{\text{SCAD-MAVE},\mathcal{A}^C} = 0$.

(b) Asymptotic normality:

$$\sqrt{n_0}\left[\hat{\beta}_{\text{SCAD-MAVE},\mathcal{A}} - \beta_{\mathcal{A}}\right] \xrightarrow{d} \text{N}(0, \Sigma_{\beta,\mathcal{A}}),$$

where $\Sigma_{\beta,\mathcal{A}} = \sigma^2 J_0^{-1}$ with $J_0 = E\left[\{X_{\mathcal{A}} m'(\beta_{\mathcal{A}}^\top X_{\mathcal{A}})\}\{X_{\mathcal{A}} m'(\beta_{\mathcal{A}}^\top X_{\mathcal{A}})\}^\top\right] - E\Big\{E\left[\{X_{\mathcal{A}} m'(\beta_{\mathcal{A}}^\top X_{\mathcal{A}})\}\right.$ $\left|\beta_{\mathcal{A}}^\top X_{\mathcal{A}} = U\right] E\left[\{X_{\mathcal{A}} m'(\beta_{\mathcal{A}}^\top X_{\mathcal{A}})\}|\beta_{\mathcal{A}}^\top X_{\mathcal{A}} = U\right]^\top\Big\}$.

The consequence of Lemma 1 is that $\hat{\beta}_{\text{SCAD-MAVE},\mathcal{A}}$ satisfies Assumption A6 with

$$\sqrt{n_0}\left(\hat{\beta}_{\text{SCAD-MAVE},\mathcal{A}} - \beta_{\mathcal{A}}\right) = \frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} \phi_{\mathcal{A}}(X_j, Y_j) + o_p(1),$$

where $\phi_{\mathcal{A}}(X_j, Y_j) = J_0^{-1}\left\{E\left[X_{\mathcal{A}} m'(\beta_{\mathcal{A}}^\top X_{\mathcal{A}})|\beta_{\mathcal{A}}^\top X_{\mathcal{A}} = U\right] - X_{\mathcal{A}} m'(\beta_{\mathcal{A}}^\top X_{\mathcal{A}})\right\}\varepsilon_j$. Denote $\delta_{a,\mathcal{A}} = E\left[m'(Z_i) X_{i,\mathcal{A}}^\top\right]$. Similar to Theorem 1, the following theorem holds true with its proof given in Appendix.

Theorem 3: Under the conditions imposed in Theorem 1 and Assumptions B1 - B4, one has

$$\sqrt{n_1}\left(\hat{\Delta}_{\text{SCAD}} - \Delta\right) \xrightarrow{d} \text{N}\left(0, \sigma^2_{\Delta,\text{SCAD}}\right),$$

where $\sigma^2_{\Delta,\text{SCAD}} = \sigma_1^2 + \lambda\left(\sigma_2^2 + \sigma^2_{3,\mathcal{A}} + 2\delta_{a,\mathcal{A}}\Sigma_{23,\mathcal{A}}\right)$, σ_1^2 and σ_2^2 defined in Theorem 1, $\sigma^2_{3,\mathcal{A}} = \delta_{a,\mathcal{A}}\Sigma_{\beta,\mathcal{A}}\delta_{a,\mathcal{A}}^\top$, $\Sigma_{\beta,\mathcal{A}} = \text{Var}(\phi_{\mathcal{A}}(X_j, Y_j))$ and $\Sigma_{23,\mathcal{A}} = \text{Cov}(r(Z_j)\varepsilon_j, \phi_{\mathcal{A}}(X_j, Y_j))$ for $j = 1, \ldots, n_0$.

3 Monte Carlo Simulation Studies

In these simulation studies, we investigate the finite sample performances for our proposed estimator, the proposed method to choose covariates, and the proposed Bootstrap procedure.

3.1 Evaluating the Proposed Estimator

In this subsection, we evaluate our proposed estimators $\hat{\Delta}$ for ATE in (5) through a series of Monte Carlo simulations. To illustrate the universality of our method, we consider both

linear and nonlinear potential models for outcomes. Notice that the SCM method implicitly requires that the predictors of the treated and control observations are positively correlated, while our method is not restricted by this positive correlation requirement. Hence, we also evaluate our method in the cases where the predictors of the treated and control observations are potentially negatively correlated. In each setting, we set the dimension of covariates as $d = 5$ and $d = 10$. When $d = 5$, the true index vector is $\beta = (1,\, 0.7,\, -0.5,\, 0.25,\, 0.8)^\top$, and $\beta = (1,\, 0.7,\, -0.5,\, 0.5,\, -0.75,\, 0.8,\, -0.4,\, 1,\, -0.2,\, 0.2)^\top$ for $d = 10$. In total, we consider the following data generating processes.

Example 1: (linear and nonlinear potential outcomes) We consider the following linear and nonlinear models for the potential outcomes, $Y(0) = m\left(\beta^\top X\right)+\varepsilon$ and $Y(1) = Y(0)+2$, where for $j = 1, \cdots, d$, $X_j \sim \mathrm{N}(0, 1)$ for the control units, and $X_j \sim \mathrm{U}(-\sqrt{3}, \sqrt{3})$ for the treated units, and $\varepsilon \sim \mathrm{N}(0, 1)$. The sample sizes are $n_0 = 100,\, 200,\, 500$ and $n_1 = 100,\, 200,\, 500$. In this DGP, we consider two cases: linear model as $m(u) = u$ and nonlinear model as $m(u) = u^2$, respectively.

In the simulation studies, we consider several choices of h. For simplicity, we present the results for only bandwidth $h = 0.5\, n_0^{-1/3}$, which satisfies clearly the assumption requirements. Here, the Gaussian kernel $K(v) = \frac{1}{\sqrt{2\pi}} \exp(-v^2/2)$ is used. For each setting, we conduct 500 Monte Carlo simulations and we compare our method with the SCM method. We can see that under each setting, the true ATE is $\Delta = 2$. For each setting, among all 500 simulations, both the root mean square error (RMSE), which is defined by $\mathrm{RMSE} = \left[\sum_{k=1}^{500} \left(\hat{\Delta}_k - \Delta\right)^2 /500\right]^{1/2}$, and the mean of 500 absolute deviation errors (MADE), which is given by $\mathrm{ADE}_k = \left|\hat{\Delta}_k - \Delta\right|$, where $\hat{\Delta}_k$ is the estimate for the kth simulation, are recorded for two estimators. The Monte Carlo simulation results are shown in Table 1 for the linear model in the top panel and nonlinear model in the bottom panel. Note that for different choices of h as long as h satisfies the assumption requirements, the finite sample performances are almost same and the results are available upon request.

Now, we discuss the performances of our method and the SCM in detail. First, one can see clearly from the results given in the top panel in Table 1 for $m(u) = u$, where the potential outcome model is linear, that both methods perform well and our method is slightly better

Table 1: Different estimators for Δ in Example 1.

$m(u) = u$								
			$n_0 = 100$		$n_0 = 200$		$n_0 = 500$	
n_1	d	method	RMSE	MADE	RMSE	MADE	RMSE	MADE
100	5	SCM	0.2059	0.1653	0.1710	0.1357	0.1475	0.1186
		QSCM	0.1600	0.1251	0.1243	0.0979	0.1099	0.0871
	10	SCM	0.2131	0.1685	0.1701	0.1354	0.1405	0.1111
		QSCM	0.1618	0.1273	0.1274	0.0983	0.1116	0.0887
200	5	SCM	0.1857	0.1505	0.1487	0.1180	0.1261	0.0991
		QSCM	0.1331	0.1069	0.1013	0.0771	0.0822	0.0664
	10	SCM	0.1967	0.1577	0.1432	0.1139	0.1158	0.0932
		QSCM	0.1368	0.1079	0.1019	0.0795	0.0837	0.0674
500	5	SCM	0.1728	0.1359	0.1393	0.1121	0.1152	0.0920
		QSCM	0.1197	0.0946	0.0847	0.0657	0.0621	0.0501
	10	SCM	0.1748	0.1360	0.1291	0.1050	0.0986	0.0773
		QSCM	0.1238	0.0971	0.0873	0.0678	0.0627	0.0506

$m(u) = u^2$								
			$n_0 = 100$		$n_0 = 200$		$n_0 = 500$	
n_1	d	method	RMSE	MADE	RMSE	MADE	RMSE	MADE
100	5	SCM	1.5115	1.3871	1.9292	1.8537	2.4095	2.3735
		QSCM	0.1747	0.1378	0.1370	0.1089	0.1102	0.0877
	10	SCM	1.3799	1.1350	1.6911	1.5168	2.2791	2.1884
		QSCM	0.2579	0.2022	0.1750	0.1335	0.1246	0.0981
200	5	SCM	1.5321	1.4126	1.9098	1.8434	2.4005	2.3664
		QSCM	0.1619	0.1264	0.1155	0.0915	0.0820	0.0664
	10	SCM	1.4186	1.1595	1.7042	1.5464	2.2657	2.1955
		QSCM	0.2388	0.1833	0.1547	0.1227	0.0948	0.0746
500	5	SCM	1.5411	1.4073	1.8588	1.7988	2.4045	2.3730
		QSCM	0.1478	0.1205	0.0978	0.0763	0.0653	0.0519
	10	SCM	1.3649	1.1149	1.6627	1.5183	2.1978	2.1430
		QSCM	0.2401	0.1860	0.1438	0.1091	0.0756	0.0611

than the SCM. Next, from the results shown in the bottom panel in Table 1 for $m(u) = u^2$, where the potential outcome model is nonlinear, it is obvious that the SCM is invalid and our method performs much better, especially with smaller sample size and higher dimension of covariates.

In summary, the finite sample performance of the proposed estimator is well-behaved in the sense that both the RMSE and MADE values are generally small. The RMSE and

MADE values decrease as no matter the sample size n_0 or n_1 increases. This is in line with our expectation in the sense that, as in Theorem 1, the asymptotic results of the proposed estimator is related to both n_0 and n_1. Clearly, the estimation is compromised by the dimension of the covariates. However, this negative effect of the dimension of the covariates on the estimation is faint in our quasi synthetic control method as expected.

3.2 Evaluating the Proposed Variable Selection Method

In this subsection, we conduct a series of Monte Carlo simulations to evaluate the effectiveness combining our QSCM estimator with the variable selection methods introduced in Section 2.6. For simplicity, we only illustrate the performance in Example 1 in Section 3.1. The dimension of covariates is set as $d = 20$ with $\beta = (1, 0.7, -0.5, 0.25, 0.8, 0, \ldots, 0)^\top$ as a 20-dimensional vector with only five nonzero entries. This choice of β indicates that only the first five covariates are predictable for the outcome variable.

As in Section 3.1, we still set the bandwidth $h = 0.5\, n_0^{-1/3}$ and use the Gaussian kernel. For each setting, the simulation is repeated 500 times. We also use the RMSE and the MADE as main evaluation metrics for the estimation of the treatment effect. We compare the QSCM estimator without variable selection in (5) and the penalized QSCM estimator. As conventionally, we evaluate the performance of variable selection by the mean of true positive rate (TPR) and false positive rate (FPR) based on 500 replications. The TPR and FPR are the most widely used evaluation metrics in the field of variable selection. Adapting to our setting, the TPR and FPR are defined as follows: $\text{TPR} = \#\{1 \le j \le 5 : \hat{\beta}_j \neq 0\}/5$ and $\text{FPR} = \#\{6 \le j \le 20 : \hat{\beta}_j \neq 0\}/15$. The top panel in Table 2 summarizes the simulation results under $m(u) = u$. We can see that the proposed variable selection method works well, since the true positive rate is close to 1 and the false positive rate is relatively small and tends to 0 as the sample size n_0 increases. After selecting relevant variables in the first step of estimating the index parameter β, the treatment effect estimator in the second step also behaves well. Both RMSE and MADE are comparably small and decrease as no matter the sample size n_0 or n_1 increases.

The simulation results under $m(u) = u^2$ are presented in the bottom panel in Table 2. The true positive rate is exactly 1 for all settings, indicating that the proposed variable

Table 2: Simulation results for the variable selection procedure.

$m(u) = u$							
		QSCM		pen-QSCM		Variable Selection	
n_0	n_1	RMSE	MADE	RMSE	MADE	TPR	FPR
100	100	0.1719	0.1365	0.1683	0.1316	0.9032	0.0980
	200	0.1437	0.1143	0.1434	0.1152	0.8892	0.0872
	500	0.1373	0.1089	0.1272	0.1003	0.8900	0.0900
200	100	0.1302	0.1041	0.1300	0.1024	0.9464	0.0453
	200	0.1074	0.0835	0.1060	0.0821	0.9412	0.0453
	500	0.0902	0.0708	0.0878	0.0682	0.9384	0.0493
500	100	0.1111	0.0871	0.1110	0.0880	0.998	0.0263
	200	0.0854	0.0683	0.0838	0.0677	0.9924	0.0223
	500	0.0634	0.0516	0.0621	0.0503	0.9936	0.0225

$m(u) = u^2$							
		QSCM		pen-QSCM		Variable Selection	
n_0	n_1	RMSE	MADE	RMSE	MADE	TPR	FPR
100	100	0.2009	0.1585	0.1887	0.1484	0.9992	0.0273
	200	0.1806	0.1428	0.1705	0.1353	0.9984	0.0303
	500	0.1563	0.1248	0.1439	0.1132	0.9988	0.0332
200	100	0.1458	0.1179	0.1391	0.1110	1.0000	0.0073
	200	0.1210	0.0937	0.1149	0.0880	1.0000	0.0089
	500	0.0961	0.0730	0.0949	0.0745	1.0000	0.0084
500	100	0.1142	0.0917	0.1105	0.0877	0.998	0.0049
	200	0.0887	0.0711	0.0859	0.0695	1.0000	0.0027
	500	0.0668	0.0536	0.0649	0.0523	1.0000	0.0035

selection method can correctly select all relevant variables and the false positive rate is tiny, indicating that the proposed variable selection method can exclude most irrelevant variables.

3.3 Evaluating the Proposed Bootstrap Procedure

In this subsection, a series of Monte Carlo simulations are conducted to evaluate the validity of the Bootstrap procedure proposed in Section 2.5. We use the same data generating process as in Section 3.1 except that now for $j = 1, \ldots, d$, $X_j \sim \mathrm{U}(-\sqrt{2}, \sqrt{2})$ for the treated units, and the Bootstrap replication is set as $B = 500$. As in Section 3.1, we still set the bandwidth $h = 0.5 n_0^{-1/3}$ and use the Gaussian kernel. For each setting, the simulation is repeated 1000 times. Given the significance level α, we evaluate the corresponding coverage rate of the confidence interval constructed by the proposed Bootstrap procedure. The

Table 3: Simulation results for inference based on the proposed Bootstrap procedure.

			$m(u) = u$			$m(u) = u^2$		
n_1	d	α	$n_0 = 100$	$n_0 = 200$	$n_0 = 500$	$n_0 = 100$	$n_0 = 200$	$n_0 = 500$
100	5	0.9	0.894	0.900	0.911	0.899	0.914	0.920
		0.95	0.950	0.954	0.954	0.954	0.959	0.957
		0.99	0.991	0.989	0.997	0.993	0.991	0.995
	10	0.9	0.906	0.900	0.903	0.924	0.926	0.912
		0.95	0.945	0.950	0.961	0.965	0.971	0.962
		0.99	0.988	0.986	0.998	0.993	0.993	0.999
200	5	0.9	0.893	0.902	0.914	0.895	0.902	0.917
		0.95	0.945	0.947	0.959	0.940	0.954	0.957
		0.99	0.987	0.992	0.992	0.986	0.992	0.992
	10	0.9	0.888	0.907	0.920	0.912	0.933	0.921
		0.95	0.932	0.955	0.958	0.959	0.960	0.967
		0.99	0.982	0.991	0.990	0.993	0.996	0.994
500	5	0.9	0.875	0.910	0.911	0.903	0.915	0.912
		0.95	0.941	0.958	0.958	0.950	0.953	0.963
		0.99	0.995	0.989	0.992	0.994	0.994	0.991
	10	0.9	0.873	0.899	0.913	0.905	0.938	0.921
		0.95	0.930	0.949	0.962	0.949	0.979	0.967
		0.99	0.986	0.995	0.997	0.992	0.996	0.999

simulation results for $\alpha = 0.10, 0.05, 0.01$ are depicted in Table 3, from which, we can see that, under different settings, the proposed Bootstrap procedure results in reasonably good estimated coverage probabilities.

4 Revisit of the NSW Data

In this section, we study an empirical application by using our quasi synthetic control method to analyze the data from the National Supported Work Demonstration. The NSW program was a labor market program for underprivileged workers operated during the mid-1970s in the United States. By providing these workers with subsidized job for 9 to 18 months, the NSW program aimed to strengthen their job skills and enhance their employment opportunities. The NSW program randomly assigned the qualified applicants to the treatment and control groups, making the program a randomized controlled trial, which is universally recognized as the golden standard to learn the treatment effect. This appealing feature of the NSW program motivates numerous researches.

Lalonde (1986) first analyzes the male sub-sample of the NSW program, which contains $n_1 = 297$ treated units and $n_0 = 425$ control units. In the Lalonde sample, the outcome of interest is the annual earnings in 1978. Additionally, the Lalonde sample also collects several individual characteristics: age, education, black, hispanic, married, no degree, and annual earnings in 1975. Based on the Lalonde sample, the average treatment effect is \$886. The Lalonde sample only collects one year of pre-treatment earnings (the annual earning in 1975). However, the existing literature has pointed out that it is necessary to include more than one year of pre-treatment earnings since many applicants for training programs experience a drop in their earnings just prior to joining the training program. Therefore, Dehejia and Wahba (1999) reorganize the Lalonde sample and collected the annual earnings in 1974. By excluding the individuals with the annual earnings in 1974 missed, the Dehejia-Wahba sample consists of $n_1 = 185$ treated units and $n_0 = 260$ control units, and the ATT estimate based on the Dehejia-Wahba sample is \$1794, termed as the experimental benchmark value[2]. Note that the Dehejia-Wahba sample has been widely used in many empirical studies. For example, Dehejia and Wahha (2002) apply the propensity score matching method to this dataset by using the the Dehejia-Wahba sample. However, as pointed out by Smith and Todd (2005), estimates of the impact of NSW based on propensity score matching are highly sensitive to the set of variables included in the propensity score model, while Abadie & Imbens (2011) evaluate the performance of various matching estimators by analyzing the NSW data. For more literature on analyzing this dataset, the reader is referred to the paper by Abadie and L'Hour (2021) and references therein.

Both the Lalonde and the Dehejia-Wahha samples are based on experimental data and provide us with two unbiased estimates of the ATE. To evaluate different estimators for treatment effects, it is recommended to use a non-experimental control group and estimate the treatment effect based on the experimental treated and non-experimental control groups. Lalonde (1986) constructs six non-experimental control groups from the Panel Study of Income Dynamics (PSID) and the Current Population Survey, as well as further subsets subtracted from these two basic control groups. Referring to the existing literature, we use

[2]For details on how to compute this benchmark value, please refer to the paper by Dehejia and Wahba (1999) or Abadie and L'Hour (2021).

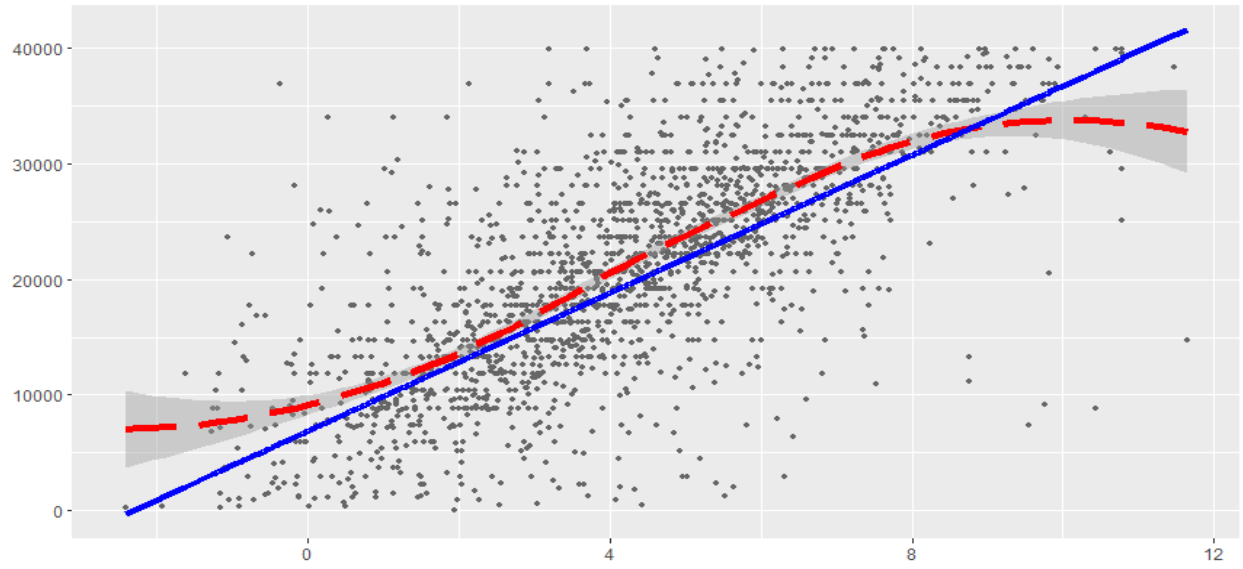

Figure 1: Scatterplot of Y_0 versus Z in PSID group, together with the lowess estimate of the unknown function $m(\cdot)$ in the dashed red line with its pointwise 95% confidence interval presented by the shaded area and a least-squares fitting of $m(\cdot)$ in the solid blue line.

the experimental treated group from the Dehejia-Wahha sample ($n_1 = 185$) and the non-experimental group from the Population Survey of Income Dynamics ($n_0 = 2490$) to illustrate our quasi synthetic control method. The outcome variable Y_i is the annual earnings in 1978 and 10 covariates in X_i are considered. Table 4 presents summary statistics for the three groups used in our analysis.

First, we would like to see if there exists a nonlinear relationship between the outcome and the index. To do so in a visual way, using data from PSID group, we plot the outcome Y_0 (y-axis) versus the estimated single index Z (x-axis) in Figure 1, together with a nonparametric estimate (*lowess* in R, locally-weighted polynomial regression technique) of the unknown function $m(\cdot)$ in the dashed red line (with its pointwise 95% confidence interval presented by the shaded area), and a least-squares fitting of $m(\cdot)$ in the solid blue line. From Figure 1, it is clear that there does exist a nonlinear relationship between Y_{0i} and Z_i and this supports strongly that our nonlinear model is appropriate for this real data.

Now, to compute the QSCM estimator $\hat{\Delta}$, as in Monte Carlo simulations, we use the Gaussian kernel and the bandwidth h is selected by an ad hoc approach, which is 0.23, in the sense that it minimizes the absolute value of the bias of the estimated ATE. Also,

Table 4: Summary statistics of 10 covariates.

	Experimental data								Non-experimental data			
	Treated ($n_1 = 185$)				Control ($n_0 = 260$)				PSID ($n_0 = 2490$)			
	Mean	Std	Min	Max	Mean	Std	Min	Max	Mean	Std	Min	Max
Covariates												
Age	25.82	7.16	17	48	25.05	7.06	17	55	34.85	10.44	18	55
Education	10.35	2.01	4	16	10.09	1.61	3	14	12.12	3.08	0	17
Black	0.84	0.36	0	1	0.83	0.38	0	1	0.25	0.43	0	1
Hispanic	0.06	0.24	0	1	0.11	0.31	0	1	0.03	0.18	0	1
Married	0.19	0.39	0	1	0.15	0.36	0	1	0.87	0.34	0	1
No degree	0.71	0.46	0	1	0.83	0.37	0	1	0.31	0.46	0	1
Earnings in 1974	2095.57	4886.62	0	35040.07	2107.03	5687.91	0	39570.68	19428.75	13406.88	0	137148.7
Earnings in 1975	1532.06	3219.25	0	25142.24	1266.91	3102.98	0	23031.98	19063.34	13596.95	0	156653.2
Unemployment in 1974	0.71	0.46	0	1	0.75	0.43	0	1	0.09	0.28	0	1
Unemployment in 1975	0.6	0.49	0	1	0.68	0.47	0	1	0.1	0.3	0	1
Outcome variable												
Earnings in 1978	6349.14	7867.4	0	60307.93	4554.8	5483.84	0	39483.53	21553.92	15555.35	0	121173.6

we compare our quasi synthetic control estimator with a series of existing estimators: the conventional synthetic control estimator (**SCM**), the penalized synthetic control estimator which minimizes the bias (**Pen. SCM**) as in Abadie and L'Hour (2021), and the one-match nearest neighbor matching estimator (**1-Matching**).[3] Table 5 reports the empirical results. These four estimators yield treatment effects ranging from \$1801.22 to \$2138.8. Given the

Table 5: Non-experimental estimates for the NSW data.

Method	Benchmark	QSCM	SCM	Pen.SCM	1-Matching
Treatment effect	1794.34	1801.22	2118.61	1881.40	2138.80

Note: The QSCM estimate is computed based on the optimal bandwidth $h = 0.23$, which minimizes the absolute value of the bias of the estimated ATE. The results for Pen SCM and 1-Matching in this table come from Abadie and L'Hour (2021).

experimental benchmark $\Delta = \$1794.34$, our QSCM estimator is best in the sense that it has the smallest bias from the target $\Delta = \$1794.34$. This result indicates that our method captures well the unknown features of the NSW data with a possible nonlinear relationship between Y_{0i} and Z_i.

Finally, it is also very interesting to notice that in this empirical example, the conventional SCM needs to optimize a 2490×1 vector of weights for each out of total 185 treated units, which is computationally expensive in practice. Our computing is carried out on a IBM X3550M4 dual processors server equipped with Twenty-four Core Intel Xeon E5-2620 v2 @ 2.10GHz CPU, 64 GB RAM running Windows Server 2019. Using parallel computing in R language, it takes us 1.69 hours to compute the conventional SCM estimate. Whereas, given a selected bandwidth, the computation time for our QSCM estimate is 13.6 seconds without parallel computation. To gauge this phenomenon, indeed, as pointed out by Abadie and and L'Hour (2021), the best synthetic control may not be unique with many treated units and/or many control units. Therefore, to search for the best synthetic control, the computing is too heavy so that our method can save a huge computing time.

[3]Here, the one-match nearest neighbor matching estimator means that, for each unit in the treated group, we find its nearest neighbor in PSID group with respect to the distance between covariates. And units in PSID group can be reused and matched to multiple treated units.

5 Conclusion

The SC method is a popular and powerful way of estimating ATE as addressed by Athey and Imbens (2017). But, as pointed in the literature, the SC methods have some shortcomings. To overcome these difficulties, this paper proposes a QSC method, which can accommodate nonlinearity and feature a fast computing. In particular, this article provides the inference theory for the QSC method and we derive the asymptotic distribution of the QSC ATE estimators with and without penalty term. Also, due to complex structure of the asymptotic variances of the proposed estimators, we resolve the difficulty by proposing a carefully designed and easy-to-implement Bootstrap method and establish the validity of subsampling method for inference. Our work complements the conventional SC method and its variants. In addition, our simulations show that the QSC method performs well in practice. Finally, we apply the QSC method to estimate ATE for the NSW data. The empirical application demonstrates that when the conventional SC method fits the data poorly, the QSC method can fit the data well and provide reasonable ATE estimation results.

Finally, it is worth to note that under the current framework, one might extend easily the proposed methodology to estimate the quantile (distributional) treatment effects as investigated in Cai et al. (2022), which is warranted as a future research topic.

References

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque country. *American Economic Review*, 93(1):113-132.

Abadie, A. and Imbens, G.W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235-267.

Abadie, A. and Imbens, G.W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1-11.

Abadie, A. and L'Hour, J. (2021). A Penalized Synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, 116(536):1817-1834.

Athey, S. and Imbens, G.W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(1):3-32.

Cai, Z., Das, M., Xiong, H. and Wu, X. (2006). Functional coefficient instrumental variables models. *Journal of Econometrics* 133(1):207-241.

Cai, Z., Fang, Y., Lin, M. and Zhan, M. (2022). Estimating quantile treatment effects for panel data. *Working Paper*, Department of Economics, University of Kansas. https://ideas.repec.org/p/kan/wpaper/202205.html.

Cai, Z., Juhl, T. and Yang, B. (2015). Functional index coefficient models with variable selection. *Journal of Econometrics*, 189(2):272-284.

Cattaneo, M.D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2):138-154.

Chiburis, R. C. (2010). Semiparametric bounds on treatment effects. *Journal of Econometrics*, 159(2):267-275.

Dehejia, R.H., and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053-1062.

Dehejia, R.H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151-161.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Fitting and Its Applications*. Chapman and Hall, New York.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348-1360.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259-276.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, 76(376), 817-823.

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, 86(1):77-90.

Galvao, A. F. and Wang, L. (2015). Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association*, 110(512):1528-1542.

Heckman, J.J., Ichimura, H. and Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies*, 65(2):261-294.

Hsiao, C., Ching, S. and Wan, S.K. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong kong with mainland China. *Journal of Applied Econometrics*, 27(5):705-740.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1-2):71-120.

Kang, J.D.Y and Schafer, J.L. (2007). Demystifying double robustness: A Comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523-539.

Kong, E. and Xia, Y. (2007). Variable selection for the single-index model. *Biometrika*, 94(1):217-229.

LaLonde, R.J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604-620.

Li, K. (2020). Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, 115(532):2068-2083.

Naik, P.A. and Tsai, C.-L. (2001). Single-index model selections. *Biometrika*, 88(3):821-832.

Otsu, T. and Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112(520):1720-1732.

Ouyang, M. and Peng, Y. (2015). The treatment-effect estimation: A case study of the 2008 economic stimulus package of China. *Journal of Econometrics*, 188(2):545-557.

Park, H., Petkova, E., Tarpey, T. and Ogden, R. T. (2021). A constrained single index regression for estimating interactions between a treatment and covariates. *Biometrics*, 77(2):506-518.

Peng, H. and Huang, T. (2011). Penalized least squares for single index models. *Journal of Statistical Planning and Inference*, 141(4):1362-1379.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688-701.

Serfling, R.J. (1980). *Approximation Theorems in Mathematical Statistics*. John Wiley & Sons, New York.

Smith, J.A. and Todd, P.E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators?. *Journal of Econometrics*, 125(1-2):305-353.

Sun, Y., Yan, K.X. and Li, Q. (2021). Estimation of average treatment effect based on a semiparametric propensity score. *Econometric Reviews*, 40(9):852-866.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B*, 58(1):267-288.

Wan, S.K., Xie, Y. and Hsiao, C. (2018). Panel data approach vs synthetic control method. *Economics Letters*, 164(C):121-123.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics & Data Analysis*, 52(9):4512-4520.

Wang, T., Xu, P. and Zhu, L. (2013). Penalized minimum average variance estimation. *Statistica Sinica*, 23(2):543-569.

Wu, Y., Ren, T. and Mu, L. (2016). Importance reweighting using adversarial-collaborative training. *Neural Information Processing Systems, 2016 Workshop on Adversarial Training*.

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, 22(6):1112-1137.

Xia, Y., Tong, H., Li, W.K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space (with discussions). *Journal of the Royal Statistical Society, Series B*, 64(2):363-410.

Zeng, P., He, T. and Zhu, Y. (2012). A LASSO-type approach for estimation and variable selection in single index models. *Journal of Computational and Graphical Statistics*, 21(1):92-109.

Zhang, H., Huang, L. and Liu, L.L. (2020). On Bootstrap consistency of MAVE for single index models. *Computational Statistics & Data Analysis*, 141(C):28-39.

Appendix: Mathematical Proofs

To establish the asymptotic theory of $\hat{\Delta}$, we first state some preliminaries, stated here for convenience without their proofs. Indeed, Lemma 2 comes from Theorem 5.4A in Serfling (1980, p.190) for the generalized U-statistic for two samples and Lemma 3 comes from Lemma 1 in Heckman, Ichimura and Todd (1998).

Lemma 2: Suppose $\{M_{0,j}\}_{j=1}^{n_0}$ and $\{M_{1,i}\}_{i=n_0+1}^{n}$ are independent and within each group and they are i.i.d. Define the generalized U-statistic,

$$U_{n_0}(\psi_{n_0,n_1}) = \frac{1}{n_0 \cdot n_1} \sum_{i=n_0+1}^{n} \sum_{j=1}^{n_0} \psi_{n_0,n_1}(M_{0,j}, M_{1,i})$$

with $E\{\psi_{n_0,n_1}(M_{0,j}, M_{1,i})\} = 0$, and its projection

$$\hat{U}_{n_0}(\psi_{n_0,n_1}) = \frac{1}{n_0} \sum_{j=1}^{n_0} p_0(M_{0,j}) + \frac{1}{n_1} \sum_{i=n_0+1}^{n} p_1(M_{1,i}),$$

where $p_0(M_{0,j}) = E\{\psi_{n_0,n_1}(M_{0,j}, M_{1,i})|M_{0,j}\}$ and $p_1(M_{1,i}) = E\{\psi_{n_0,n_1}(M_{0,j}, M_{1,i})|M_{1,i}\}$, respectively. If $0 < \lim n_1/n_0 = \lambda < \infty$, where $n = n_0 + n_1$ and $E\{\psi_{n_0,n_1}(M_{0,j}, M_{1,i})^2\} = o(n_0) + o(n_1)$, then,

$$nE\left[(U_{n_0}\psi_{n_0,n_1} - \hat{U}_{n_0}\psi_{n_0,n_1})^2\right] = o(1),$$

which is an extension of Theorem 5.4A in Serfling (1980, p.191) for two samples.

Lemma 3: Under our setting, consider the function $g(z) = E(Y_{0i}|Z_i = z)$. Suppose that

$$\hat{z} - z = \frac{1}{n_0} \sum_{j=1}^{n_0} \psi_{n_0}(X_j, Y_j) + o_p(n_0^{-1/2}) \quad \text{and} \quad \hat{g}(z) - g(z) = \frac{1}{n_0} \sum_{j=1}^{n_0} \omega_{n_0}(X_j, Y_j) + o_p(n_0^{-1/2})$$

Also, suppose that $\partial\hat{g}(z)/\partial z$ and $\hat{z}$ are uniformly consistent and converge to $\partial g(z)/\partial z$ and z, respectively and $\partial g(z)/\partial z$ is continuous. Then, we have

$$\hat{g}(\hat{z}) - g(z) = \frac{1}{n_0} \sum_{j=1}^{n_0} [\omega_{n_0}(X_j, Y_j) + \partial g(z)/\partial z \cdot \psi_{n_0}(X_j, Y_j)] + o_p(n_0^{-1/2}).$$

Proof of Theorem 1: If $\{Z_j\}_{j=1}^{n_0}$ and $\{Z_i\}_{i=n_0+1}^{n}$ are known, we can estimate $E(Y_{0i}|X_i)$ for

$i = n_0 + 1, \ldots, n$ as

$$\tilde{m}(Z_i) = \frac{\sum_{j=1}^{n_0} K_h(Z_j - Z_i)Y_j}{\sum_{l=1}^{n_0} K_h(Z_l - Z_i)} = \frac{\frac{1}{n_0}\sum_{j=1}^{n_0} K_h(Z_j - Z_i)Y_j}{\tilde{f}_p(Z_i)}, \tag{A.1}$$

where $\tilde{f}_p(u)$ denotes the kernel density estimator of $f_p(u)$ using the pre-treatment indexes $\{Z_j\}_{j=1}^{n_0}$. However, we can only use $\{\hat{Z}_j\}_{j=1}^{n_0}$ and $\{\hat{Z}_i\}_{i=n_0+1}^{n}$ to estimate $E(Y_{0i}|X_i)$ for $i = n_0 + 1, \ldots, n$ as

$$\hat{m}(\hat{Z}_i) = \frac{\sum_{j=1}^{n_0} K_h(\hat{Z}_j - \hat{Z}_i)Y_j}{\sum_{l=1}^{n_0} K_h(\hat{Z}_l - \hat{Z}_i)} = \frac{\frac{1}{n_0}\sum_{j=1}^{n_0} K_h(\hat{Z}_j - \hat{Z}_i)Y_j}{\hat{f}_p(\hat{Z}_i)}, \tag{A.2}$$

where $\hat{f}_p(u)$ denotes the kernel density estimator of $f_p(u)$ using the pre-treatment estimators $\{\hat{Z}_j\}_{j=1}^{n_0}$. By (A.2) we can write $\hat{\Delta}$ as

$$\hat{\Delta} = \frac{1}{n_1}\sum_{i=n_0+1}^{n}\left[Y_{1i} - \hat{m}(\hat{Z}_i)\right], \tag{A.3}$$

and

$$\sqrt{n_1}\left(\hat{\Delta} - \Delta\right) = \frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n}\left[\Delta_i - \Delta - \varepsilon_i\right] - \frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n}\left[\hat{m}(\hat{Z}_i) - m(Z_i)\right] = I_1 - I_2, \tag{A.4}$$

where I_1 and I_2 are well defined. Next, we discuss the asymptotics of I_2. For $i = n_0+1, \ldots, n$, denote $v_i = (Z_1, \ldots, Z_{n_0}, Z_i)^\top$ and $\hat{v}_i = (\hat{Z}_1, \ldots, \hat{Z}_{n_0}, \hat{Z}_i)^\top$, and define $g(z_1, \ldots, z_{n_0}, z_i) = \sum_{j=1}^{n_0} K_h(z_j - z_i)Y_j \left[\sum_{l=1}^{n_0} K_h(z_l - z_i)\right]^{-1}$. Thus, $\tilde{m}(Z_i) = g(v_i)$ and $\hat{m}(\hat{Z}_i) = g(\hat{v}_i)$. Again, notice that

$$\hat{v}_i - v_i = \begin{pmatrix} X_1^\top \\ \vdots \\ X_{n_0}^\top \\ X_i^\top \end{pmatrix}\left(\hat{\beta} - \beta\right) = \frac{1}{n_0}\sum_{l=1}^{n_0} A_i^\top \phi(X_l, Y_l) + o(n_0^{-1/2}), \tag{A.5}$$

where $A_i^\top = (X_1^\top, \ldots, X_{n_0}^\top, X_i^\top)$. For any $i = n_0 + 1, \ldots, n$, $Y_j = m(Z_j) + \varepsilon_j = m(Z_i) + (m(Z_j) - m(Z_i)) + \varepsilon_j$, which implies that

$$\begin{aligned}\frac{1}{n_0}\sum_{j=1}^{n_0} K_h(Z_j - Z_i)Y_j &= \frac{1}{n_0}\sum_{j=1}^{n_0} K_h(Z_j - Z_i)\left[m(Z_i) + (m(Z_j) - m(Z_i)) + \varepsilon_j\right] \\ &= \tilde{f}_p(Z_i)m(Z_i) + \frac{1}{n_0}\sum_{j=1}^{n_0} K_h(Z_j - Z_i)\left[m(Z_j) - m(Z_i) + \varepsilon_j\right]. \end{aligned} \tag{A.6}$$

Combining (A.1) and (A.6) leads to

$$\tilde{m}(Z_i) = m(Z_i) + \frac{\frac{1}{n_0}\sum_{j=1}^{n_0} K_h(Z_j - Z_i)\,[m(Z_j) - m(Z_i) + \varepsilon_j]}{\tilde{f}_p(Z_i)}.$$

Then,

$$g(v_i) - E(Y_{0i}|X_i) = \tilde{m}(Z_i) - m(Z_i) = \frac{1}{n_0}\sum_{j=1}^{n_0} \frac{K_h(Z_j - Z_i)\,[m(Z_j) - m(Z_i) + \varepsilon_j]}{\frac{1}{n_0}\sum_{l=1}^{n_0} K_h(Z_l - Z_i)}. \quad \text{(A.7)}$$

Again, by combining (A.5) and (A.7), and using Lemma 3, we have

$$\begin{aligned}
\hat{m}(\hat{Z}_i) - m(Z_i) &= \frac{1}{n_0}\sum_{j=1}^{n_0}\left\{\frac{K_h(Z_j - Z_i)\,[m(Z_j) - m(Z_i) + \varepsilon_j]}{\frac{1}{n_0}\sum_{l=1}^{n_0} K_h(Z_l - Z_i)} + m'(Z_i)X_i^\top \phi(X_j, Y_j)\right\} \\
&= \frac{\frac{1}{n_0}\sum_{j=1}^{n_0} K_h(Z_j - Z_i)\,[m(Z_j) - m(Z_i) + \varepsilon_j]}{\frac{1}{n_0}\sum_{l=1}^{n_0} K_h(Z_l - Z_i)} + m'(Z_i)X_i^\top \frac{1}{n_0}\sum_{j=1}^{n_0}\phi(X_j, Y_j) \\
&= \frac{\frac{1}{n_0}\sum_{j=1}^{n_0} K_h(Z_j - Z_i)\,[m(Z_j) - m(Z_i) + \varepsilon_j]}{\tilde{f}_p(Z_i)} + m'(Z_i)X_i^\top \frac{1}{n_0}\sum_{j=1}^{n_0}\phi(X_j, Y_j).
\end{aligned}$$

Hence,

$$\begin{aligned}
I_2 = &\frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n} \frac{\frac{1}{n_0}\sum_{j=1}^{n_0} K_h(Z_j - Z_i)\,[m(Z_j) - m(Z_i) + \varepsilon_j]}{\tilde{f}_p(Z_i)} \\
&+ \frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n} m'(Z_i)X_i^\top \frac{1}{n_0}\sum_{j=1}^{n_0}\phi(X_j, Y_j) = I_{21} + I_{22},
\end{aligned}$$

where both I_{21} and I_{22} are defined in an obvious manner. Now, I_{21} is re-expressed as follows

$$\begin{aligned}
I_{21} &= \frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n} \frac{\frac{1}{n_0}\sum_{j=1}^{n_0} K_h(Z_j - Z_i)\,[m(Z_j) - m(Z_i) + \varepsilon_j]}{\tilde{f}_p(Z_i)} \\
&= \sqrt{n_1}\,\frac{1}{n_0 \cdot n_1}\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0} \frac{K_h(Z_j - Z_i)\,[m(Z_j) - m(Z_i) + \varepsilon_j]}{\tilde{f}_p(Z_i)} \\
&= \sqrt{n_1}\,\frac{1}{n_0 \cdot n_1}\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0} \psi_{n_0,n_1}(Y_j, Z_j, Z_i),
\end{aligned}$$

where $\psi_{n_0,n_1}(Y_j, Z_j, Z_i) = K_h(Z_j - Z_i)\left[m(Z_j) - m(Z_i) + \varepsilon_j\right]/\tilde{f}_p(Z_i)$. Notice that

$$\begin{aligned} E\left[\psi_{n_0,n_1}(Y_j, Z_j, Z_i)\right] &= E\left\{E\left[\psi_{n_0,n_1}(Y_j, Z_j, Z_i)\,\middle|\, Z_i\right]\right\} \\ &= E\left\{E\left\{\frac{K_h(Z_j - Z_i)\left[m(Z_j) - m(Z_i)\right]}{\tilde{f}_p(Z_i)}\,\middle|\, Z_i\right\}\right\}, \end{aligned}$$

and

$$\begin{aligned} & E\left\{\frac{K_h(Z_j - Z_i)\left[m(Z_j) - m(Z_i)\right]}{\tilde{f}_p(Z_i)}\,\middle|\, Z_i\right\} \\ =& \frac{1}{h}\int \frac{K(\frac{z - Z_i}{h})\left[m(z) - m(Z_i)\right]}{\tilde{f}_p(Z_i)} f_p(z)\mathrm{d}z \\ =& \frac{1}{f_p(Z_i)}\left\{\int K(t)\left[m(Z_i + ht) - m(Z_i)\right] f_p(Z_i + ht)\mathrm{d}t\right\} \cdot \left[1 + O_p(h^2) + O_p\left(\frac{1}{\sqrt{n_0 h}}\right)\right] \\ =& \frac{1}{f_p(Z_i)}\left\{\int K(t)\left[htm'(Z_i) + \frac{h^2t^2}{2}m''(Z_i)\right]\left[f_p(Z_i) + htf_p'(Z_i)\right]\mathrm{d}t\right\} \\ & \times \left[1 + O_p(h^2) + O_p\left(\frac{1}{\sqrt{n_0 h}}\right)\right] = O_p(h^2), \end{aligned}$$

where $m''(z)$ is the second order derivative of $m(z)$ and $f_p'(z)$ is the first order derivative of $f_p(z)$. Clearly $E\left[\psi_{n_0,n_1}(Y_j, Z_j, Z_i)\,\middle|\, Z_i\right] = O(h^2)$ and $E\left[\psi_{n_0,n_1}(Y_j, Z_j, Z_i)\right] = O(h^2)$. Now, define $\tilde{\psi}_{n_0,n_1}(Y_j,\ Z_j,\ Z_i) = \psi_{n_0,n_1}(Y_j, Z_j, Z_i) - E\left[\psi_{n_0,n_1}(Y_j, Z_j, Z_i)\right]$. Thus, $E\left[\tilde{\psi}_{n_0,n_1}(Y_j,\ Z_j,\ Z_i)\,\middle|\, Z_i\right] = O(h^2)$ and $E\left[\tilde{\psi}_{n_0,n_1}(Y_j,\ Z_j,\ Z_i)\right] = 0$. Meanwhile,

$$\begin{aligned} E\left[\tilde{\psi}_{n_0,n_1}(Y_j,\ Z_j,\ Z_i)\,\middle|\, Y_j, Z_j\right] &= E\left\{\psi_{n_0,n_1}(Y_j, Z_j, Z_i) - E\left[\psi_{n_0,n_1}(Y_j, Z_j, Z_i)\right]\,\middle|\, Y_j,\ Z_j\right\} \\ &= E\left\{\psi_{n_0,n_1}(Y_j, Z_j, Z_i)\,\middle|\, Y_j,\ Z_j\right\} + O_p(h^2). \end{aligned}$$

Notice that

$$\begin{aligned} & E\left\{\psi_{n_0,n_1}(Y_j, Z_j, Z_i)\,\middle|\, Y_j,\ Z_j\right\} = \frac{1}{h}\int \frac{K(\frac{Z_j - z}{h})\left[m(Z_j) - m(z) + \varepsilon_j\right]}{\tilde{f}_p(z)} f_a(z)\mathrm{d}z \\ =& \left\{\int K(t)\left[m(Z_j) - m(Z_j + ht) + \varepsilon_j\right] r(Z_j + ht)\mathrm{d}t\right\} \cdot \left[1 + O_p(h^2) + O_p\left(\frac{1}{\sqrt{n_0 h}}\right)\right] \\ =& \left\{\int K(t)\left[m(Z_j) - m(Z_j + ht)\right] r(Z_j + ht)\mathrm{d}t + \varepsilon_j \int K(t) r(Z_j + ht)\mathrm{d}t\right\} \\ & \times \left[1 + O_p(h^2) + O_p\left(\frac{1}{\sqrt{n_0 h}}\right)\right]. \end{aligned}$$

It is obvious that

$$\begin{aligned}&\int K(t)\left[m(Z_j+ht)-m(Z_j)\right]r(Z_j+ht)\mathrm{d}t\\&=\int K(t)\left[htm'(Z_j)+\frac{h^2t^2}{2}m''(Z_j)\right]\left[r(Z_j)+htr'(Z_j)\right]\mathrm{d}t=O_p(h^2).\end{aligned}$$

Also,

$$\int K(t)r(Z_j+ht)\mathrm{d}t=\int K(t)\left[r(Z_j)+htr'(Z_j)+\frac{h^2t^2}{2}r''(Z_j)+o(h^2t^2)\right]\mathrm{d}t=r(Z_j)+O_p(h^2),$$

where $r''(z)$ is the second order derivative of $r(z)$. Then,

$$E\left\{\psi_{n_0,n_1}(Y_j,Z_j,Z_i)\,\middle|\,Y_j,\,Z_j\right\}=O_p(h^2)+\left\{r(Z_j)\left[1+O_p(h^2)+O_p\left(\frac{1}{\sqrt{n_0h}}\right)\right]+O_p(h^2)\right\}\varepsilon_j,$$

and

$$E\left[\tilde{\psi}_{n_0,n_1}(Y_j,\,Z_j,\,Z_i)\,\middle|\,Y_j,\,Z_j\right]=\left\{r(Z_j)\left[1+O_p(h^2)+O_p\left(\frac{1}{\sqrt{n_0h}}\right)\right]+O(h^2)\right\}\varepsilon_j+O_p(h^2).$$

Next, I_{21} is rewritten as

$$\begin{aligned}I_{21}&=\sqrt{n_1}\frac{1}{n_0\cdot n_1}\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0}\psi_{n_0,n_1}(Y_j,\,Z_j,\,Z_i)\\&=\sqrt{n_1}\frac{1}{n_0\cdot n_1}\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0}\left\{\tilde{\psi}_{n_0,n_1}(Y_j,\,Z_j,\,Z_i)+E\left[\psi_{n_0,n_1}(Y_j,\,Z_j,\,Z_i)\right]\right\}\\&=\sqrt{n_1}\frac{1}{n_0\cdot n_1}\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0}\tilde{\psi}_{n_0,n_1}(Y_j,\,Z_j,\,Z_i)+o_p(1)\\&=\tilde{I}_{21}+o_p(1).\end{aligned}$$

It follows from Lemma 2 that $\tilde{I}_{21}$ is asymptotically equivalent to

$$\begin{aligned}&\sqrt{n_1}\left\{\frac{1}{n_0}\sum_{j=1}^{n_0}E\left[\tilde{\psi}_{n_0,n_1}(Y_j,\,Z_j,\,Z_i)\,\middle|\,Y_j,\,Z_j\right]+\frac{1}{n_1}\sum_{i=n_0+1}^{n}E\left[\tilde{\psi}_{n_0,n_1}(Y_j,\,Z_j,\,Z_i)\,\middle|\,Z_i\right]\right\}\\=&\sqrt{\frac{n_1}{n_0}}\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\left\{r(Z_j)\left[1+O_p(h^2)+O_p\left(\frac{1}{\sqrt{n_0h}}\right)\right]+O(h^2)\right\}\varepsilon_j+o_p(1)\\=&\sqrt{\frac{n_1}{n_0}}\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}r(Z_j)\varepsilon_j+o_p(1),\end{aligned}$$

so that

$$I_{21} = \sqrt{\frac{n_1}{n_0}} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} r(Z_j)\varepsilon_j + o_p(1). \tag{A.8}$$

Now, we rewrite I_{22} as

$$\begin{aligned} I_{22} &= \sqrt{\frac{n_1}{n_0}} \left[\frac{1}{n_1} \sum_{i=n_0+1}^{n} m'(Z_i)X_i^\top\right] \left[\frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} \phi(X_j, Y_j)\right] \\ &= \sqrt{\frac{n_1}{n_0}} \left[\delta_a + o_p(1)\right] \left[\frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} \phi(X_j, Y_j)\right] = \sqrt{\frac{n_1}{n_0}} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} \delta_a\phi(X_j, Y_j) + o_p(1), \end{aligned} \tag{A.9}$$

where $\delta_a = E\left[m'(Z_i)X_i^\top\right]$. By (A.8) and (A.9), we have

$$I_2 = \sqrt{\lambda} \frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} \left[r(Z_j)\varepsilon_j + \delta_a\phi(X_j, Y_j)\right] + o_p(1) \tag{A.10}$$

Based on (A.4) and (A.10), we have

$$\sqrt{n_1}\left(\hat{\Delta} - \Delta\right) = I_1 - \sqrt{\lambda}\frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} \left[r(Z_j)\varepsilon_j + \delta_a\phi(X_j, Y_j)\right] + o_p(1),$$

from which, it is easy to see that I_1 and the second term on the right hand side is independent. Obviously, by the central limit theorem, $I_1 \xrightarrow{d} \mathrm{N}(0, \sigma_1^2)$, where $\sigma_1^2 = \mathrm{Var}(\Delta_i) + \mathrm{Var}(\varepsilon_i)) - 2\mathrm{Cov}(\Delta_i, \varepsilon_i)$ given in Theorem 1. Again, it follows by the central limit theorem that $\frac{1}{\sqrt{n_0}} \sum_{j=1}^{n_0} \left[r(Z_j)\varepsilon_j + \delta_a\phi(X_j, Y_j)\right]$ converges to a normal distribution with mean 0 and variance $\sigma_2^2 + \sigma_3^2 + 2\delta_a\Sigma_{23}$, where σ_2^2, σ_3^2, and Σ_{23} are given in Theorem 1. Therefore,

$$\sqrt{n_1}\left(\hat{\Delta} - \Delta\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma_\Delta^2\right),$$

where $\sigma_\Delta^2 = \sigma_1^2 + \lambda\left(\sigma_2^2 + \sigma_3^2 + 2\delta_a\Sigma_{23}\right)$ given in Theorem 1. This completes the proof of Theorem 1.

Proof of Theorem 2: Before embracing on the proof of Theorem 2, let P denote the distribution of $\{X_l, Y_l\}_{l=1}^n$ and use P^* to denote the Bootstrap distribution, which is the distribution of $\{X_j, Y_j^*\}_{j=1}^{n_0}$ and $\{X_i^*, Y_i^*\}_{i=n_0+1}^n$ conditional on $\{X_j, Y_j\}_{j=1}^{n_0}$ and $\{X_i, Y_i\}_{i=n_0+1}^n$. Also, we use E^* to denote the expectation with respect to P^*. Furthermore, let $S_1, S_2 \ldots$ be a sequence of random variables and $a_1, a_2, \ldots$ be a sequence of positive real numbers. Define $S_n = o_p^*(a_n)$ if for any $\varepsilon > 0$ and $\epsilon > 0$, $\lim_{n\to\infty} P\{P^*(|S_n/a_n| > \epsilon) > \varepsilon\} = 0$. Similarly, $S_n = O_p^*(a_n)$ means that if for any $\varepsilon > 0$ and $\epsilon > 0$, there exists $M > 0$ such that

$\limsup_{n\to\infty} P\{P^*(|S_n/a_n| > M) > \varepsilon\} < \epsilon$.

Notice that

$$\begin{aligned}
\sqrt{n_1}(\hat{\Delta}^* - \hat{\Delta}) &= \frac{1}{\sqrt{n_1}} \sum_{i=n_0+1}^{n} \left\{ \left[Y_i^* - \hat{m}^* \left((X_i^*)^\top \hat{\beta}^* \right) \right] - \left[\frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i - \frac{1}{n_1} \sum_{i=n_0+1}^{n} \hat{m}(X_i^\top \hat{\beta}) \right] \right\} \\
&= \frac{1}{\sqrt{n_1}} \sum_{i=n_0+1}^{n} \left\{ \left(Y_i^* - \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i \right) - \left[\hat{m}^* \left((X_i^*)^\top \hat{\beta}^* \right) - \frac{1}{n_1} \sum_{i=n_0+1}^{n} \hat{m}(X_i^\top \hat{\beta}) \right] \right\} \\
&= \frac{1}{\sqrt{n_1}} \sum_{i=n_0+1}^{n} \left\{ \left(Y_i^* - \frac{1}{n_1} \sum_{i=n_0+1}^{n} Y_i \right) - \left[\hat{m} \left((X_i^*)^\top \hat{\beta} \right) - \frac{1}{n_1} \sum_{i=n_0+1}^{n} \hat{m}(X_i^\top \hat{\beta}) \right] \right\} \\
&\quad - \frac{1}{\sqrt{n_1}} \sum_{i=n_0+1}^{n} \left[\hat{m}^* \left((X_i^*)^\top \hat{\beta}^* \right) - \hat{m} \left((X_i^*)^\top \hat{\beta} \right) \right] \\
&= I_1^* - I_2^*,
\end{aligned}$$

where both I_1^* and I_2^* are well defined and

$$\hat{m}^*((X_i^*)^\top \hat{\beta}^*) = \frac{\sum_{j=1}^{n_0} K_h \left(X_j^\top \hat{\beta}^* - (X_i^*)^\top \hat{\beta}^* \right) Y_j^*}{\sum_{j=1}^{n_0} K_h \left(X_j^\top \hat{\beta}^* - (X_i^*)^\top \hat{\beta}^* \right)} = \frac{\frac{1}{n_0} \sum_{j=1}^{n_0} K_h \left(X_j^\top \hat{\beta}^* - (X_i^*)^\top \hat{\beta}^* \right) Y_j^*}{\hat{f}_p^* \left((X_i^*)^\top \hat{\beta}^* \right)}$$

with $\hat{f}_p^*(u)$ denoting the kernel density estimator of $f_p(u)$ using the pre-treatment estimators $\{X_j^\top \hat{\beta}^*\}_{j=1}^{n_0}$. Next, we discuss the term I_2^*. To apply Lemma 3, define

$$\tilde{m}^* \left((X_i^*)^\top \hat{\beta} \right) = \frac{\sum_{j=1}^{n_0} K_h \left(X_j^\top \hat{\beta} - (X_i^*)^\top \hat{\beta} \right) Y_j^*}{\sum_{j=1}^{n_0} K_h \left(X_j^\top \hat{\beta} - (X_i^*)^\top \hat{\beta} \right)} = \frac{\frac{1}{n_0} \sum_{j=1}^{n_0} K_h \left(X_j^\top \hat{\beta} - (X_i^*)^\top \hat{\beta} \right) Y_j^*}{\hat{f}_p \left((X_i^*)^\top \hat{\beta} \right)}. \tag{A.11}$$

Also, denote $\hat{v}_i = \left(X_1^\top \hat{\beta}, \ldots, X_{n_0}^\top \hat{\beta}, (X_i^*)^\top \hat{\beta} \right)^\top$ and $\hat{v}_i^* = \left(X_1^\top \hat{\beta}^*, \ldots, X_{n_0}^\top \hat{\beta}^*, (X_i^*)^\top \hat{\beta}^* \right)^\top$ for $i = n_0 + 1, \ldots, n$, and define

$$g(z_1, \ldots, z_{n_0}, z_i) = \frac{\sum_{j=1}^{n_0} K_h(z_j - z_i) Y_j^*}{\sum_{j=1}^{n_0} K_h(z_j - z_i)}.$$

Thus, $\tilde{m}^* \left((X_i^*)^\top \hat{\beta} \right) = g(\hat{v}_i)$ and $\hat{m}^* \left((X_i^*)^\top \hat{\beta}^* \right) = g(\hat{v}_i^*)$. Furthermore, by following the

proof of Theorem 1 in Zhang, Huang and Liu (2020), one has

$$\hat{v}_i^* - \hat{v}_i = \begin{pmatrix} X_1^\top \\ \vdots \\ X_{n_0}^\top \\ (X_i^*)^\top \end{pmatrix} \left(\hat{\beta}^* - \hat{\beta}\right) = A_i^\top \left[\frac{1}{n_0}\sum_{j=1}^{n_0} \phi(X_j, Y_j)\xi_j + o_p(n_0^{-1/2})\right]$$
$$= \frac{1}{n_0}\sum_{j=1}^{n_0} A_i^\top \phi(X_j, Y_j)\xi_j + o_p(n_0^{-1/2}), \tag{A.12}$$

where $A_i^\top = \left(X_1^\top, \ldots, X_{n_0}^\top, (X_i^*)^\top\right)$ and $\phi(X_j, Y_j)$ is the same as in Assumption A6. For any $i = n_0 + 1, \ldots, n$,

$$Y_j^* = \hat{Y}_j + \varepsilon_j^* = \hat{m}\left((X_i^*)^\top\hat{\beta}\right) + \left[\hat{m}(X_j^\top\hat{\beta}) - \hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right] + \varepsilon_j^*,$$

which implies that

$$\begin{aligned}
&\frac{1}{n_0}\sum_{j=1}^{n_0} K_h\left(X_j^\top\hat{\beta} - (X_i^*)^\top\hat{\beta}\right) Y_j^* \\
=&\frac{1}{n_0}\sum_{j=1}^{n_0} K_h\left(X_j^\top\hat{\beta} - (X_i^*)^\top\hat{\beta}\right)\left\{\hat{m}\left((X_i^*)^\top\hat{\beta}\right) + \left[\hat{m}(X_j^\top\hat{\beta}) - \hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right] + \varepsilon_j^*\right\} \\
=&\hat{f}_p\left((X_i^*)^\top\hat{\beta}\right)\hat{m}\left((X_i^*)^\top\hat{\beta}\right) + \frac{1}{n_0}\sum_{j=1}^{n_0} K_h\left(X_j^\top\hat{\beta} - (X_i^*)^\top\hat{\beta}\right)\left\{\left[\hat{m}(X_j^\top\hat{\beta}) - \hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right] + \varepsilon_j^*\right\}.
\end{aligned} \tag{A.13}$$

Combining (A.11), (A.12) and (A.13) leads to

$$\tilde{m}^*\left((X_i^*)^\top\hat{\beta}\right) = \hat{m}\left((X_i^*)^\top\hat{\beta}\right) + \frac{\frac{1}{n_0}\sum_{j=1}^{n_0} K_h\left(X_j^\top\hat{\beta} - (X_i^*)^\top\hat{\beta}\right)\left\{\left[\hat{m}(X_j^\top\hat{\beta}) - \hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right] + \varepsilon_j^*\right\}}{\hat{f}_p\left((X_i^*)^\top\hat{\beta}\right)},$$

which means that

$$g(\hat{v}_i) - \hat{m}(\hat{Z}_i) = \frac{1}{n_0}\sum_{j=1}^{n_0} \frac{K_h\left(X_j^\top\hat{\beta} - (X_i^*)^\top\hat{\beta}\right)\left\{\left[\hat{m}(X_j^\top\hat{\beta}) - \hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right] + \varepsilon_j^*\right\}}{\frac{1}{n_0}\sum_{l=1}^{n_0} K_h\left(X_l^\top\hat{\beta} - (X_i^*)^\top\hat{\beta}\right)}. \tag{A.14}$$

A combination of (A.11) - (A.14), together with Lemma 3, implies that

$$\begin{aligned}\hat{m}^*\left((X_i^*)^\top\hat{\beta}^*\right)-\hat{m}\left((X_i^*)^\top\hat{\beta}\right)=&\frac{1}{n_0}\sum_{j=1}^{n_0}\left\{\frac{K_h\left(X_j^\top\hat{\beta}-(X_i^*)^\top\hat{\beta}\right)\left\{\left[\hat{m}(X_j^\top\hat{\beta})-\hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right]+\varepsilon_j^*\right\}}{\frac{1}{n_0}\sum_{l=1}^{n_0}K_h\left(X_l^\top\hat{\beta}-(X_i^*)^\top\hat{\beta}\right)}\right.\\&\left.+\frac{\partial\hat{m}\left((X_i^*)^\top\hat{\beta}\right)}{\partial\hat{v}_i}A_i^\top\phi(X_j,Y_j)\xi_j\right\}\\=&\frac{\frac{1}{n_0}\sum_{j=1}^{n_0}K_h\left(X_j^\top\hat{\beta}-(X_i^*)^\top\hat{\beta}\right)\left\{\left[\hat{m}(X_j^\top\hat{\beta})-\hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right]+\varepsilon_j^*\right\}}{\hat{f}_p\left((X_i^*)^\top\hat{\beta}\right)}\\&+\frac{\partial\hat{m}\left((X_i^*)^\top\hat{\beta}\right)}{\partial\hat{v}_i}A_i^\top\frac{1}{n_0}\sum_{j=1}^{n_0}\phi(X_j,Y_j)\xi_j.\end{aligned}$$

Hence,

$$\begin{aligned}I_2^*=&\frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n}\left[\hat{m}^*\left((X_i^*)^\top\hat{\beta}^*\right)-\hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right]\\=&\frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n}\frac{\frac{1}{n_0}\sum_{j=1}^{n_0}K_h\left(X_j^\top\hat{\beta}-(X_i^*)^\top\hat{\beta}\right)\left\{\left[\hat{m}(X_j^\top\hat{\beta})-\hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right]+\varepsilon_j^*\right\}}{\hat{f}_p\left((X_i^*)^\top\hat{\beta}\right)}\\&+\frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n}\frac{\partial\hat{m}\left((X_i^*)^\top\hat{\beta}\right)}{\partial\hat{v}_i}A_i^\top\frac{1}{n_0}\sum_{j=1}^{n_0}\phi(X_j,Y_j)\xi_j=I_{21}^*+I_{22}^*,\end{aligned}$$

where both I_{21}^* and I_{22}^* are defined in an obvious manner. First, we consider I_{21}^*. To this end, I_{21}^* is re-expressed as

$$\begin{aligned}I_{21}^*=&\frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n}\frac{\frac{1}{n_0}\sum_{j=1}^{n_0}K_h\left(X_j^\top\hat{\beta}-(X_i^*)^\top\hat{\beta}\right)\left\{\left[\hat{m}(X_j^\top\hat{\beta})-\hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right]+\varepsilon_j^*\right\}}{\hat{f}_p\left((X_i^*)^\top\hat{\beta}\right)}\\=&\sqrt{n_1}\frac{1}{n_0\cdot n_1}\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0}\frac{K_h\left(X_j^\top\hat{\beta}-(X_i^*)^\top\hat{\beta}\right)\left\{\left[\hat{m}(X_j^\top\hat{\beta})-\hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right]+\varepsilon_j^*\right\}}{\hat{f}_p\left((X_i^*)^\top\hat{\beta}\right)}.\end{aligned}$$

Easily, one can show that

$$\sqrt{n_1}\frac{1}{n_0\cdot n_1}\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0}\frac{K_h\left(X_j^\top\hat{\beta}-(X_i^*)^\top\hat{\beta}\right)\left\{\left[\hat{m}(X_j^\top\hat{\beta})-\hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right]\right\}}{\hat{f}_p\left((X_i^*)^\top\hat{\beta}\right)}=o_p^*(1).$$

Then,

$$I_{21}^* = \sqrt{n_1}\frac{1}{n_0 \cdot n_1}\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0}\frac{K_h\left(X_j^\top\hat{\beta} - (X_i^*)^\top\hat{\beta}\right)\varepsilon_j^*}{\hat{f}_p\left((X_i^*)^\top\hat{\beta}\right)} + o_p^*(1)$$
$$= \sqrt{n_1}\frac{1}{n_0 \cdot n_1}\sum_{i=n_0+1}^{n}\sum_{j=1}^{n_0}\psi_{n_0,n_1}(Y_j^*, (X_i^*)^\top\hat{\beta}) + o_p^*(1) = \tilde{I}_{21}^* + o_p^*(1)$$

where $\psi_{n_0,n_1}(Y_j^*, (X_i^*)^\top\hat{\beta}) = K_h\left(X_j^\top\hat{\beta} - (X_i^*)^\top\hat{\beta}\right)\varepsilon_j^*/\hat{f}_p\left((X_i^*)^\top\hat{\beta}\right)$.

Obviously,

$$E^*\left[\psi_{n_0,n_1}(Y_j^*, (X_i^*)^\top\hat{\beta})\right] = 0 \quad \text{and} \quad E^*\left[\psi_{n_0,n_1}(Y_j^*, (X_i^*)^\top\hat{\beta})\,\middle|\,(X_i^*)^\top\hat{\beta}\right] = 0.$$

Meanwhile, following the same arguments as in the proof of Theorem 1, we can show that

$$E^*\left[\psi_{n_0,n_1}\left(Y_j^*, (X_i^*)^\top\hat{\beta}\right)\,\middle|\,Y_j^*\right] = O_p^*(h^2) + \left\{r(X_j^\top\hat{\beta})\left[1 + O_p^*(h^2) + O_p^*\left(\frac{1}{\sqrt{n_0 h}}\right)\right] + O_p^*(h^2)\right\}\varepsilon_j^*.$$

Then, it follows from Lemma 2 that $\tilde{I}_{21}^*$ is asymptotically equivalent to

$$\sqrt{n_1}\frac{1}{n_0}\sum_{j=1}^{n_0}E^*\left[\tilde{\psi}_{n_0,n_1}\left(Y_j^*, X_j^\top\hat{\beta}, (X_i^*)^\top\hat{\beta}\right)\,\middle|\,Y_j^*,\, X_j^\top\hat{\beta}\right]$$
$$= \sqrt{\frac{n_1}{n_0}}\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\left\{r(X_j^\top\hat{\beta})\left[1 + O_p^*(h^2) + O_p^*\left(\frac{1}{\sqrt{n_0 h}}\right)\right] + O^*(h^2)\right\}\varepsilon_j^* + o_p^*(1)$$
$$= \sqrt{\frac{n_1}{n_0}}\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}r(X_j^\top\hat{\beta})\varepsilon_j^* + o_p^*(1),$$

so that

$$I_{21}^* = \sqrt{\frac{n_1}{n_0}}\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}r(\hat{Z}_j)\varepsilon_j^* + o_p^*(1). \tag{A.15}$$

By Taylor expansion, we have

$$r(\hat{Z}_j) = r(Z_j) + r'(Z_j)X_j^\top(\hat{\beta} - \beta) + \frac{1}{2}r''(\tilde{Z}_j)(\hat{\beta} - \beta)^\top X_j X_j^\top(\hat{\beta} - \beta),$$

which implies that

$$\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}r(\hat{Z}_j)\varepsilon_j^* = \frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}r(Z_j)\varepsilon_j^* + \frac{1}{n_0}\sum_{j=1}^{n_0}r'(Z_j)\varepsilon_j^* X_j^\top\sqrt{n_0}(\hat{\beta} - \beta) + R_{n_0},$$

where

$$R_{n_0} = \frac{1}{2}\sqrt{n_0}(\hat{\beta}-\beta)^\top \frac{1}{n_0^{3/2}} \sum_{j=1}^{n_0} r''(\tilde{Z}_j)\varepsilon_j^* X_j X_j^\top \sqrt{n_0}(\hat{\beta}-\beta).$$

It can be shown easily that $R_{n_0} = o_p^*(1)$ and $\frac{1}{n_0}\sum_{j=1}^{n_0} r'(Z_j)\varepsilon_j^* X_j^\top \sqrt{n_0}(\hat{\beta}-\beta) = o_p^*(1)$. Hence, (A.15) becomes to

$$I_{21}^* = \sqrt{\frac{n_1}{n_0}}\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0} r(Z_j)\varepsilon_j^* + o_p^*(1) \tag{A.16}$$

Now, we consider I_{22}^*. Notice that

$$I_{22}^* = \sqrt{\frac{n_1}{n_0}}\left[\frac{1}{n_1}\sum_{i=n_0+1}^{n} \frac{\partial \hat{m}\left((X_i^*)^\top\hat{\beta}\right)}{\partial \hat{v}_i} A_i^\top\right]\left[\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\phi(X_j,Y_j)\xi_j\right].$$

Then, as $n_1 \to \infty$, it is easy to see that

$$\begin{aligned}\frac{1}{n_1}\sum_{i=n_0+1}^{n} \frac{\partial \hat{m}\left((X_i^*)^\top\hat{\beta}\right)}{\partial \hat{v}_i} A_i^\top &= \frac{1}{n_1}\sum_{i=n_0+1}^{n} \frac{\partial m\left((X_i^*)^\top\hat{\beta}\right)}{\partial \hat{v}_i} A_i^\top + o_p^*(1)\\ &= \frac{1}{n_1}\sum_{i=n_0+1}^{n} m'\left((X_i^*)^\top\hat{\beta}\right)(X_i^*)^\top + o_p^*(1).\end{aligned}$$

Then,

$$I_{22}^* = \sqrt{\frac{n_1}{n_0}}\left[\frac{1}{n_1}\sum_{i=n_0+1}^{n} m'\left((X_i^*)^\top\hat{\beta}\right)(X_i^*)^\top + o_p^*(1)\right]\left[\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\phi(X_j,Y_j)\xi_j\right]$$

Again, notice that $\frac{1}{n_1}\sum_{i=n_0+1}^{n} m'\left((X_i^*)^\top\hat{\beta}\right)(X_i^*)^\top = E^*\left[m'\left((X_i^*)^\top\hat{\beta}\right)(X_i^*)^\top\right] + o_p^*(1)$, and

$$E^*\left[m'\left((X_i^*)^\top\hat{\beta}\right)(X_i^*)^\top\right] = \frac{1}{n}\sum_{i=n_0+1}^{n} m'(\hat{Z}_i)X_i^\top = E\left[m'(Z_i)X_i^\top\right] + o_p(1).$$

Hence,

$$I_{22}^* = \sqrt{\frac{n_1}{n_0}}\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\delta_a\phi(X_j,Y_j)\xi_j + o_p^*(1), \tag{A.17}$$

where δ_a is defined in Theorem 1. Combining (A.16) and (A.17), we have

$$I_2^* = \sqrt{\lambda}\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\left[r(Z_j)\varepsilon_j^* + \delta_a\phi(X_j,Y_j)\xi_j\right] + o_p^*(1).$$

Therefore,

$$\sqrt{n_1}\left(\hat{\Delta}^* - \hat{\Delta}\right) = I_1^* - \sqrt{\lambda}\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\left[r(Z_j)\varepsilon_j^* + \delta_a\phi(X_j, Y_j)\xi_j\right] + o_p^*(1),$$

from which, it is easy to see that I_1^* and the second term on the right hand side is independent. Obviously, conditional on the original sample $\{X_i,\, Y_i\}_{i=n_0+1}^n$,

$$E^*(Y_i^*) = \frac{1}{n_1}\sum_{i=n_0+1}^{n} Y_i \quad \text{and} \quad E^*\left[\hat{m}\left((X_i^*)^\top\hat{\beta}\right)\right] = \frac{1}{n_1}\sum_{i=n_0+1}^{n}\hat{m}(X_i^\top\hat{\beta}).$$

Then, it follows from the central limit theorem that, conditional on the original sample $\{X_i,\, Y_i\}_{i=n_0+1}^n$, $I_1^* \xrightarrow{d} \mathrm{N}(0, \sigma_1^2)$, where

$$\sigma_1^2 = \lim_{n_1\to\infty}\mathrm{Var}\left(Y_i - \hat{m}(X_i^\top\hat{\beta})\right) = \mathrm{Var}(Y_{1i}) + \mathrm{Var}\left(m(Z_i)\right) - 2\mathrm{Cov}\left(Y_{1i}, m(Z_i)\right).$$

Again, by the central limit theorem, conditional on the original sample $\{X_j,\, Y_j\}_{j=1}^{n_0}$,

$$\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\left[r(Z_j)\varepsilon_j^* + \delta_a\phi(X_j, Y_j)\xi_j\right] \xrightarrow{d} \mathrm{N}\left(0, \sigma_2^2 + \sigma_3^2 + 2\delta_a\Sigma_{23}\right),$$

where $\sigma_2^2 = \lim_{n_0\to\infty}\mathrm{Var}\left(r(Z_j\varepsilon_j^*)\right) = \mathrm{Var}\left(r(Z_j)\varepsilon_j\right)$, $\sigma_3^2 = \lim_{n_0\to\infty}\mathrm{Var}\left(\delta_a\phi(X_j, Y_j)\xi_j\right) = \delta_a^\top\Sigma_\beta\delta_a$, and $\Sigma_{23} = \lim_{n_0\to\infty}\mathrm{Cov}\left(\phi(X_j, Y_j)\xi_j, r(Z_j)\varepsilon_j^*\right) = \mathrm{Cov}\left(\phi(X_j, Y_j), r(Z_j)\varepsilon_j\right)$. That is to say, conditional on the original sample $\{X_j,\, Y_j\}_{j=1}^{n_0}$ and $\{X_i,\, Y_i\}_{i=n_0+1}^n$,

$$\sqrt{n_1}\left(\hat{\Delta}^* - \hat{\Delta}\right) \xrightarrow{d} \mathrm{N}(0, \sigma_\Delta^2),$$

where σ_Δ^2 is defined in Theorem 1. This establishes the proof of Theorem 2.

Proof of Theorem 3: Following the arguments in the proof of Theorem 1, we have

$$\sqrt{n_1}\left(\hat{\Delta}_{\mathrm{SCAD}} - \Delta\right) = \frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n}\left\{\Delta_i - \Delta - \varepsilon_i\right\} - \frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n}\left[\hat{m}(\hat{Z}_i) - m(Z_i)\right] = J_1 - J_2,$$

where $J_1 \xrightarrow{d} \mathrm{N}(0, \sigma_1^2)$ with σ_1^2 defined in Theorem 1. Furthermore,

$$\begin{aligned} J_2 = &\frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n}\frac{\frac{1}{n_0}\sum_{j=1}^{n_0}K_h(Z_j - Z_i)\left[m(Z_j) - m(Z_i) + \varepsilon_j\right]}{\tilde{f}_p(Z_i)} \\ &+ \frac{1}{\sqrt{n_1}}\sum_{i=n_0+1}^{n} m'(Z_i)X_{i,\mathcal{A}}^\top\frac{1}{n_0}\sum_{j=1}^{n_0}\phi_{\mathcal{A}}(X_j, Y_j) = J_{21} + J_{22}, \end{aligned}$$

where $J_{21} = \sqrt{\lambda}\,\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0} r(Z_j)\varepsilon_j + o_p(1)$ and $J_{22} = \sqrt{\lambda}\,\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0} \delta_{a,\mathcal{A}}\phi_{\mathcal{A}}(X_j, Y_j) + o_p(1)$ with $\delta_{a,\mathcal{A}}$ defined in Theorem 3. Then,

$$J_2 = \sqrt{\lambda}\,\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\left[r(Z_j)\varepsilon_j + \delta_{a,\mathcal{A}}\phi_{\mathcal{A}}(X_j, Y_j)\right] + o_p(1) \xrightarrow{d} \mathrm{N}\left(0, \sigma_2^2 + \sigma_{3,\mathcal{A}}^2 + 2\delta_{a,\mathcal{A}}\Sigma_{23,\mathcal{A}}\right),$$

where σ_2^2, $\sigma_{3,\mathcal{A}}^2$, $\delta_{a,\mathcal{A}}$, and $\Sigma_{23,\mathcal{A}}$ are given in Theorem 3. Hence,

$$\sqrt{n_1}\left(\hat{\Delta}_{\mathrm{SCAD}} - \Delta\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma_{\Delta,\mathrm{SCAD}}^2\right),$$

where $\sigma_{\Delta,\mathrm{SCAD}}^2$ is defined in Theorem 3. This concludes the proof of Theorem 3.